



Artificial Intelligence
Index Report 2023

CHAPTER 2: Technical Performance

CHAPTER 2 PREVIEW:

Technical Performance

Overview	4
Chapter Highlights	5
2.1 What's New in 2022: A Timeline	6
2.2 Computer Vision—Image	13
Image Classification	13
ImageNet	13
Face Detection and Recognition	14
National Institute of Standards and Technology Face Recognition Vendor Test (FRVT)	15
Deepfake Detection	16
Celeb-DF	16
Human Pose Estimation	17
MPII	17
Semantic Segmentation	18
Cityscapes Challenge, Pixel-Level Semantic Labeling Task	18
Medical Image Segmentation	19
Kvasir-SEG	19
Object Detection	20
Common Objects in Context (COCO)	20
Image Generation	21
CIFAR-10 and STL-10	21

Narrative Highlight: A Closer Look at Progress in Image Generation	22
Visual Reasoning	24
Visual Question Answering (VQA) Challenge	24
Narrative Highlight: The Rise of Capable Multimodal Reasoning Systems	25
Visual Commonsense Reasoning (VCR)	27
2.3 Computer Vision—Video	28
Activity Recognition	28
Kinetics-400, Kinetics-600, Kinetics-700	28
Narrative Highlight: A Closer Look at the Progress of Video Generation	30
2.4 Language	31
English Language Understanding	31
SuperGLUE	31
Reading Comprehension Dataset Requiring Logical Reasoning (ReClor)	32
Narrative Highlight: Just How Much Better Have Language Models Become?	34
Narrative Highlight: Planning and Reasoning in Large Language Models	35
Text Summarization	36
arXiv and PubMed	36



CHAPTER 2 PREVIEW (CONT'D): Technical Performance

Natural Language Inference	37
Abductive Natural Language Inference (aNLI)	37
Sentiment Analysis	38
SST-5 Fine-Grained Classification	38
Multitask Language Understanding	39
Massive Multitask Language Understanding (MMLU)	39
Machine Translation (MT)	40
Number of Commercially Available MT Systems	40
2.5 Speech	41
Speech Recognition	41
VoxCeleb	41
Narrative Highlight:	
Whisper	42
2.6 Reinforcement Learning	44
Reinforcement Learning Environments	44
Procgen	44
Narrative Highlight:	
Benchmark Saturation	46

2.7 Hardware	47
MLPerf Training Time	47
MLPerf Inference	49
Trends in GPUs	50
2.8 Environment	52
Environmental Impact of Select Large Language Models	52
Narrative Highlight: Using AI to Optimize Energy Usage	54
2.9 AI for Science	55
Accelerating Fusion Science Through Learned Plasma Control	55
Discovering Novel Algorithms for Matrix Manipulation With AlphaTensor	55
Designing Arithmetic Circuits With Deep Reinforcement Learning	56
Unlocking de Novo Antibody Design With Generative AI	56
Appendix	57

ACCESS THE PUBLIC DATA



Overview

This year's technical performance chapter features analysis of the technical progress in AI during 2022. Building on previous reports, this chapter chronicles advancement in computer vision, language, speech, reinforcement learning, and hardware. Moreover, this year this chapter features an analysis on the environmental impact of AI, a discussion of the ways in which AI has furthered scientific progress, and a timeline-style overview of some of the most significant recent AI developments.

Chapter Highlights

Performance saturation on traditional benchmarks.

AI continued to post state-of-the-art results, but year-over-year improvement on many benchmarks continues to be marginal. Moreover, the speed at which benchmark saturation is being reached is increasing. However, new, more comprehensive benchmarking suites such as BIG-bench and HELM are being released.

AI systems become more flexible.

Traditionally AI systems have performed well on narrow tasks but have struggled across broader tasks. Recently released models challenge that trend; BEiT-3, PaLI, and Gato, among others, are single AI systems increasingly capable of navigating multiple tasks (for example, vision, language).

AI is both helping and harming the environment.

New research suggests that AI systems can have serious environmental impacts. According to Luccioni et al., 2022, BLOOM's training run emitted 25 times more carbon than a single air traveler on a one-way trip from New York to San Francisco. Still, new reinforcement learning models like BCOOLER show that AI systems can be used to optimize energy usage.

Generative AI breaks into the public consciousness.

2022 saw the release of text-to-image models like DALL-E 2 and Stable Diffusion, text-to-video systems like Make-A-Video, and chatbots like ChatGPT. Still, these systems can be prone to hallucination, confidently outputting incoherent or untrue responses, making it hard to rely on them for critical applications.

Capable language models still struggle with reasoning.

Language models continued to improve their generative capabilities, but new research suggests that they still struggle with complex planning tasks.

The world's best new scientist ... AI?

AI models are starting to rapidly accelerate scientific progress and in 2022 were used to aid hydrogen fusion, improve the efficiency of matrix manipulation, and generate new antibodies.

AI starts to build better AI.

Nvidia used an AI reinforcement learning agent to improve the design of the chips that power AI systems. Similarly, Google recently used one of its language models, PaLM, to suggest ways to improve the very same model. Self-improving AI learning will accelerate AI progress.



The technical performance chapter begins with an overview of some of the most significant technical developments in AI during 2022, as selected by the AI Index Steering Committee.

2.1 What's New in 2022: A Timeline

Feb. 2, 2022

DeepMind Releases AlphaCode

AlphaCode, an AI system that writes computer programs at a competitive level, achieves a rank within the top 54% of participants in a human programming competition. This represents an improvement on the more complex problem-solving tasks with which AI has traditionally struggled.



Figure 2.1.1

Feb. 16, 2022

DeepMind Trains Reinforcement Learning Agent to Control Nuclear Fusion Plasma in a Tokamak

Nuclear fusion is a potential source of clean, limitless energy, but producing such energy in tokamaks is difficult due to a lack of experimental data. DeepMind simulated optimal tokamak management, an example of how AI can accelerate science and combat climate change.

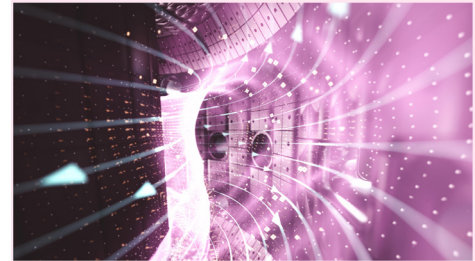


Figure 2.1.2

March 10, 2022

IndicNLG Benchmarks Natural Language Generation for Indic Languages

An international research collective launches IndicNLG, a collection of datasets for benchmarking natural language generation for 11 Indic languages. The creation of IndicNLG increases the potential for AI systems to generate language in more diverse, non-English linguistic settings.

Task	Languages	Communicative Intent	Input Type	Size
Biography Generation	L-{gu, mr}	One-sentence biography	key-value pairs	57K
Headline Generation	L	News article headlines	news article	1.31M
Sentence Summarization	L	Synonymous compact sentence	sentence	431K
Paraphrase Generation	L	Synonymous sentence	sentence	5.57M
Question Generation	L	Question leading to answer given context	context-answer pairs	1.08M

Figure 2.1.3

March 24, 2022

Meta AI Releases Make-A-Scene

Make-A-Scene is a text-to-image AI model that enables users to generate images through text. Make-A-Scene is one of many text-to-image models released in 2022.



Figure 2.1.4

April 5, 2022

Google Releases PaLM

Google's AI team trains one of the world's largest language models, PaLM. Made up of 540 billion parameters, PaLM reinforces the belief that researchers can improve performance on large language models by simply training them on more data.

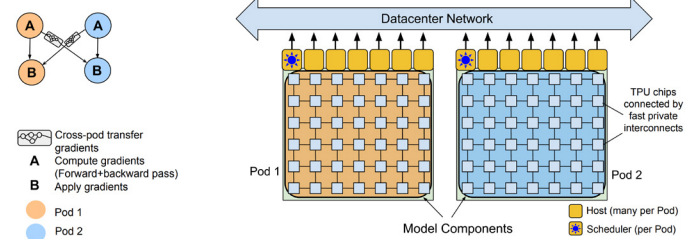


Figure 2.1.5

April 13, 2022

OpenAI Releases DALL-E 2

DALL-E 2, a text-to-image AI system that can create realistic art and images from textual descriptions, is released to the public, igniting a generative AI craze.

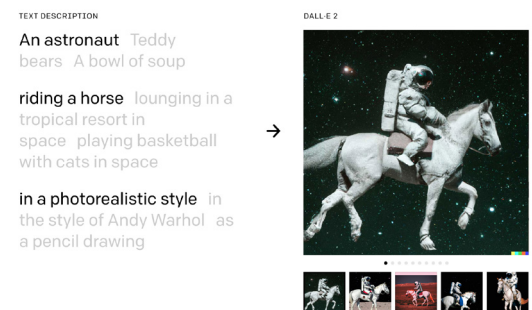


Figure 2.1.6

May 12, 2022

DeepMind Launches Gato

Gato is a new reinforcement learning agent capable of doing a wide range of tasks such as robotic manipulation, game playing, image captioning, and natural language generation. The release of such models suggests that AI systems are becoming better at generalization.



Figure 2.1.7

May 23, 2022

Google Releases Imagen

Imagen is a text-to-image diffusion model capable of producing images with a high degree of photorealism. Imagen's launch also comes with the release of DrawBench, a challenging new benchmark for text-to-image systems.

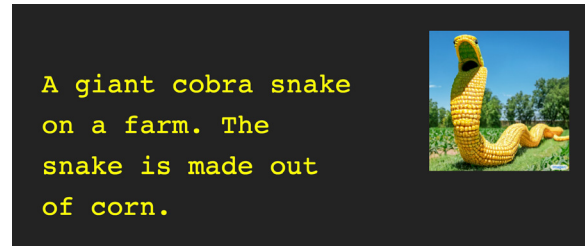


Figure 2.1.8

June 9, 2022

442 Authors Across 132 Institutions Team Up to Launch BIG-bench

In order to better challenge increasingly capable large language models, a team of 442 authors across 132 institutions launch the Beyond the Imitation Game benchmark (BIG-bench). The benchmark consists of 204 tasks ranging from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, and software development.

auto_debugging	known_unknowns	parsinlu_reading_comprehension
bbq_lite_json	language_identification	play_dialog_same_or_different
code_line_description	linguistics_puzzles	repeat_copy_logic
conceptual_combinations	logic_grid_puzzle	strange_stories
conlang_translation	logical_deduction	strategyqa
emoji_movie	misconceptions_russian	symbol_interpretation
formal_fallacies_...	novel_concepts	vitaminc_fact_verification
hindu_knowledge	operators	winowhy

Figure 2.1.9

June 21, 2022

GitHub Makes Copilot Available as a Subscription-Based Service for Individual Developers

Copilot is a generative AI system capable of turning natural language prompts into coding suggestions across multiple languages. Similar systems include OpenAI's Codex and Salesforce's CodeGen. Surveys suggest that Copilot makes coders more productive and less frustrated.

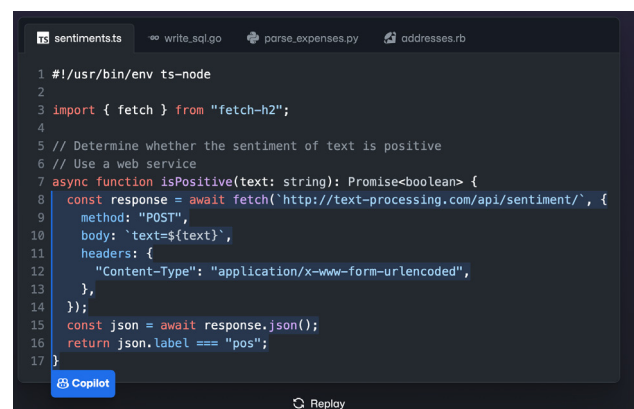


Figure 2.1.10

July 8, 2022

Nvidia Uses Reinforcement Learning to Design Better-Performing GPUs

Nvidia uses its AI systems to improve the performance of its latest H100 class of GPU chips. GPUs being essential to AI training, this is one example of how AI is starting to develop better AI.

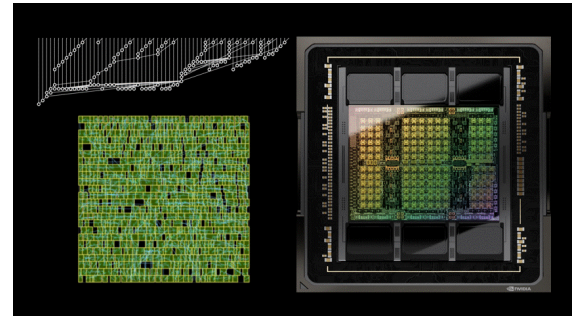


Figure 2.1.11

July 11, 2022

Meta Announces 'No Language Left Behind'

No Language Left Behind (NLLB) is a family of models that can translate across 200 distinct languages. NLLB is one of the first systems that can perform well across a wide range of low-resource languages like Kamba and Lao.

Swedish and Lingala speaker count compared with their respective published Wikipedia pages

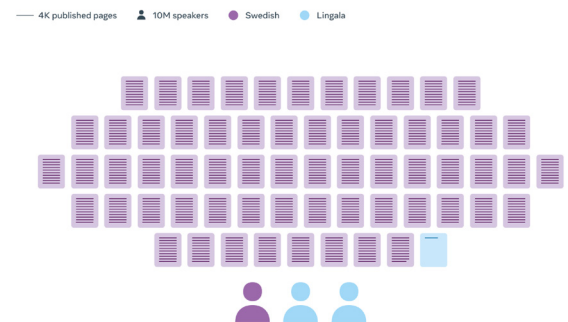


Figure 2.1.12

Aug 4, 2022

Tsinghua Researchers Launch GLM-130B

Chinese researchers affiliated with Tsinghua University release GLM-130B, a large language model that outperforms others such as Meta's OPT, Hugging Face's BLOOM, and OpenAI's original GPT-3.



Figure 2.1.13

Aug 22, 2022

Stability AI Releases Stable Diffusion

Stable Diffusion is an open-source text-to-image diffusion-based model, meaning users can freely use the model weights to generate their own images. Stable Diffusion is trained on existing images created by humans and gives no credit or acknowledgment, leaving open questions around the ethical use of image generators.



Figure 2.1.14

Sept 21, 2022

OpenAI Launches Whisper

Whisper is a large-scale speech-recognition system trained on roughly 700,000 hours of audio data and capable of respectable performance on various speech recognition tasks. The fact that Whisper required neither supervised pre-training nor unsupervised training with fine-tuning yet was able to achieve strong performance by merely increasing training data further validates the approach of increasingly scaling AI models.

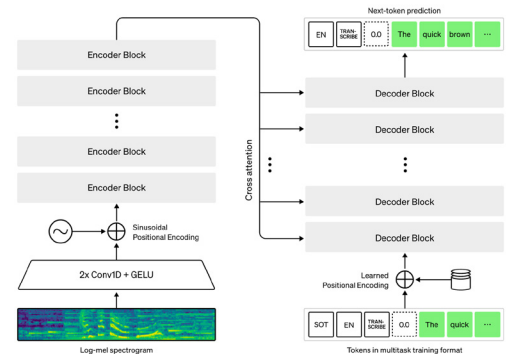


Figure 2.1.15

Sept 29, 2022

Meta Releases Make-A-Video

Make-A-Video is a system that allows users to create videos from short text descriptions. The quality of the videos is high and again demonstrates the validity of the scaling approach.



Figure 2.1.16

Oct 5, 2022

DeepMind Launches AlphaTensor

AlphaTensor is an AI reinforcement-learning-based system able to discover new and efficient algorithms for matrix manipulation. Matrix manipulation is essential to a wide range of digital practices and is a process that researchers have been trying to make more efficient for decades.

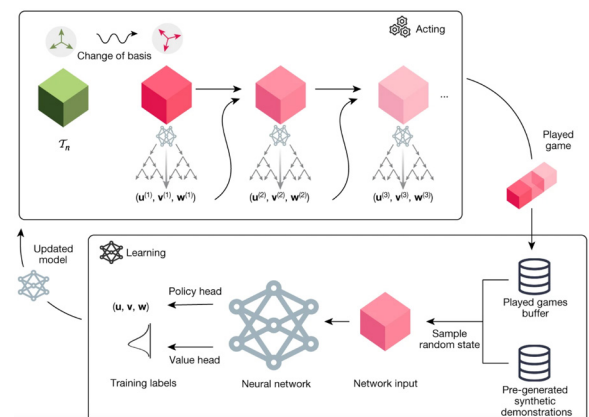


Figure 2.1.17

Oct 20, 2022

Google Uses PaLM to Improve the Reasoning of PaLM

Google researchers use one of their existing language models, PaLM, to improve the reasoning of the very same model. This process is yet another example of AI systems using their own knowledge to improve.

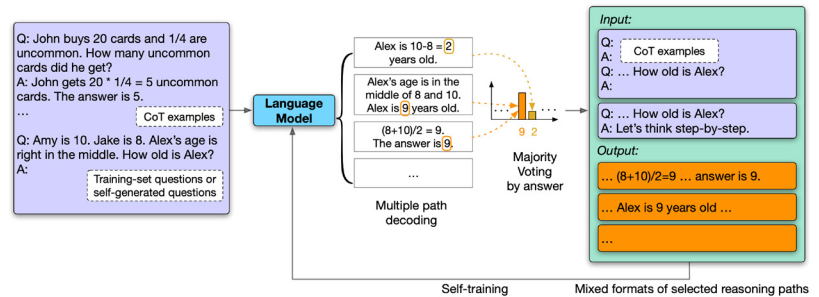


Figure 2.1.18

Nov 9, 2022

International Research Group Releases BLOOM

A collaboration of over 100 researchers from across the globe develop an open-access language model called BLOOM. BLOOM impresses with its public release and for furthering the possibilities of international collaboration in AI research.



Figure 2.1.19

Nov 16, 2022

Stanford Researchers Release HELM

As part of an attempt to judge new language models according to more unified standards, Stanford researchers develop a new benchmarking approach for large language models called Holistic Evaluation of Language Models (HELM). The launch of HELM is evidence of the AI community's attempt to develop transparency around increasingly powerful, capable, and influential large language models.

Previous work		HELM						
Scenarios	Metric	Metrics						
		Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
	Natural Questions	✓ (Accuracy)						
	XSUM	✓ (Accuracy)						
	AdversarialQA	✓ (Robustness)						
	RealToxicity Prompts	✓ (Toxicity)						
Scenarios		Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
	RAFT	✓	✓	✓	✓	✓	✓	✓
	IMDB	✓	✓	✓	✓	✓	✓	✓
	Natural Questions	✓	✓	✓	✓	✓	✓	✓
	QuAC	✓	✓	✓	✓	✓	✓	✓
	XSUM	✓				✓	✓	✓

Figure 2.1.20

Nov 22, 2022

Meta Releases CICERO

CICERO is the first AI to play in the top 10% of human participants in the game Diplomacy. CICERO's launch shows that AI systems have improved in strategic reasoning, a domain in which they have traditionally struggled, and are capable of effectively convincing humans to go along with their objectives.

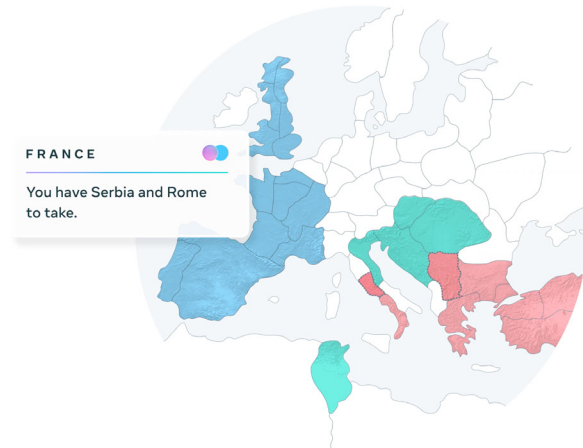


Figure 2.1.21

Nov 30, 2022

OpenAI Launches ChatGPT

ChatGPT is an impressive, publicly usable chatbot capable of writing university-level essays. Months after launching, ChatGPT reaches 100 million monthly active users, making it the fastest-growing consumer application in history. ChatGPT's release caps a year in which generative AI became a part of the zeitgeist, and raises questions about the effect that AI will have on the future of humanity.

NM

Explain to me why ChatGPT is significant?



ChatGPT is significant because it is a highly advanced AI language model developed by OpenAI, which can generate human-like text responses to questions and prompts. Its large-scale training on diverse text data and its cutting-edge deep learning architecture make it capable of generating informative and coherent responses to a wide range of topics, making it useful for various NLP applications such as chatbots, content generation, and language translation. Additionally, its open-source availability allows for further research and development in the field of AI language processing.



Figure 2.1.22

Computer vision is the subfield of AI that teaches machines to understand images and videos. Computer vision technologies have a variety of important real-world applications, such as autonomous driving, crowd surveillance, sports analytics, and video-game creation. This section tracks progress in computer vision across several different task domains which include: (1) image classification, (2) face detection and recognition, (3) deepfake detection, (4) human pose estimation, (5) semantic segmentation, (6) medical image segmentation, (7) object detection, (8) image generation, and (9) visual reasoning.

2.2 Computer Vision—Image

Image Classification

Image classification is the ability of machines to categorize objects in images (Figure 2.2.1).

ImageNet

ImageNet is one of the most widely used benchmarks for image classification. This dataset includes over 14 million images across 20,000 different object categories such as “strawberry” or “balloon.” Performance on ImageNet is measured through various accuracy metrics. Top-1 accuracy measures the degree to which the top prediction generated by an image classification model for a given image actually matches the image’s label.

As of 2022, the best image classification system on ImageNet has a top-1 accuracy rate of 91.0% (Figure 2.2.2). Although the current image classification capabilities of state-of-the-art systems is 27.7 percentage points better than a decade ago, last year saw a very marginal 0.1 percentage point improvement in classification accuracy.

A Demonstration of Image Classification

Source: Krizhevsky et al., 2012

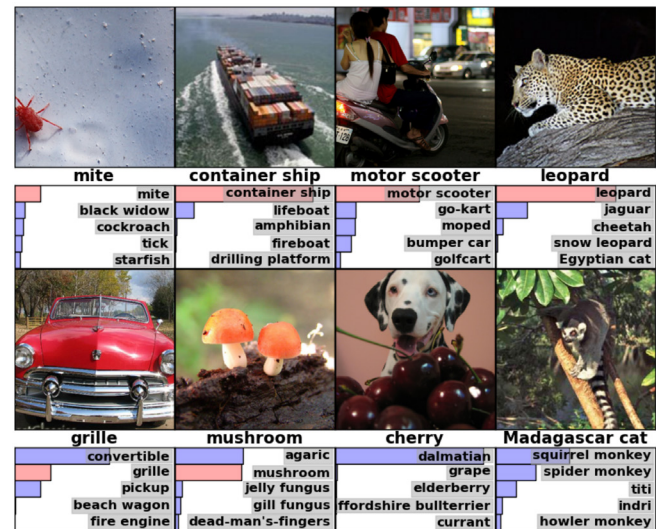


Figure 2.2.1

ImageNet Challenge: Top-1 Accuracy

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report

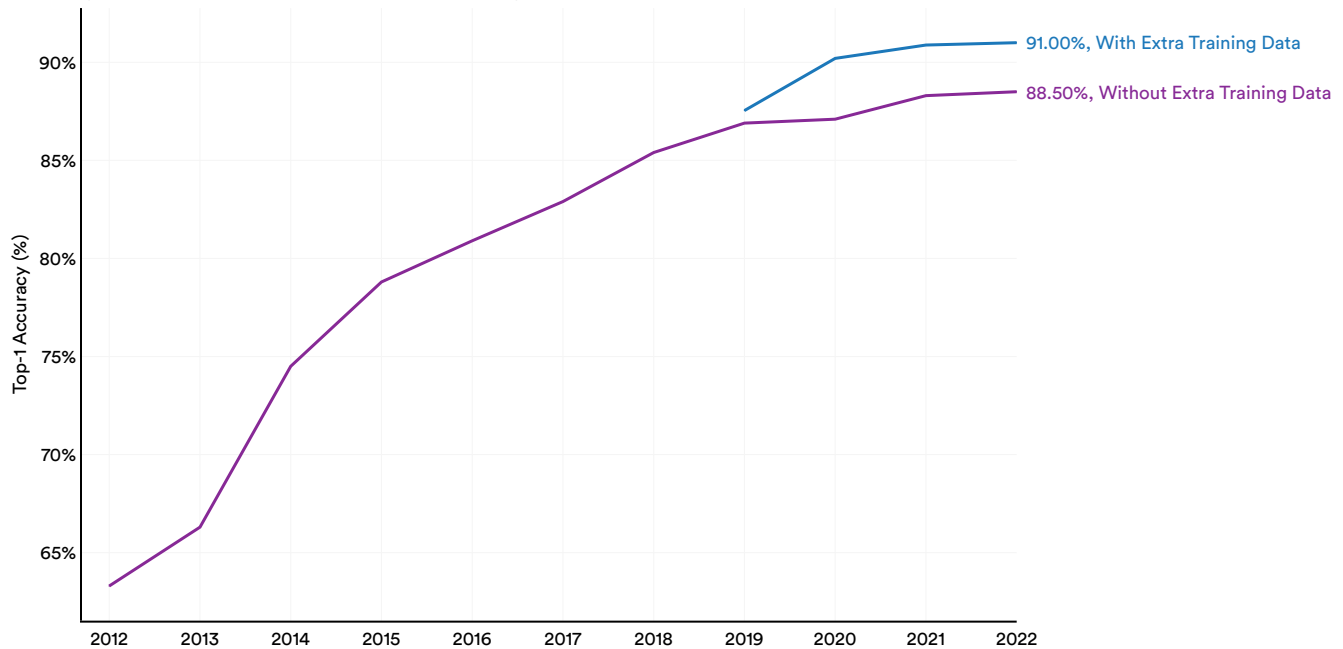


Figure 2.2.2

Face Detection and Recognition

Facial detection and recognition is the ability of AI systems to identify faces or individuals in images or videos (Figure 2.2.3). Currently, many facial recognition systems are able to successfully identify close to 100% of faces, even on challenging datasets (Figure 2.2.4).

A Demonstration of Face Detection and Recognition

Source: [Forbes, 2020](#)

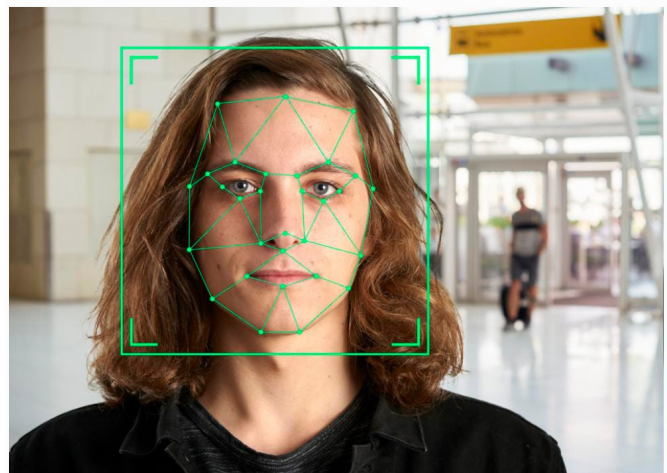


Figure 2.2.3



National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT): Verification Accuracy by Dataset

Source: National Institute of Standards and Technology, 2022 | Chart: 2023 AI Index Report

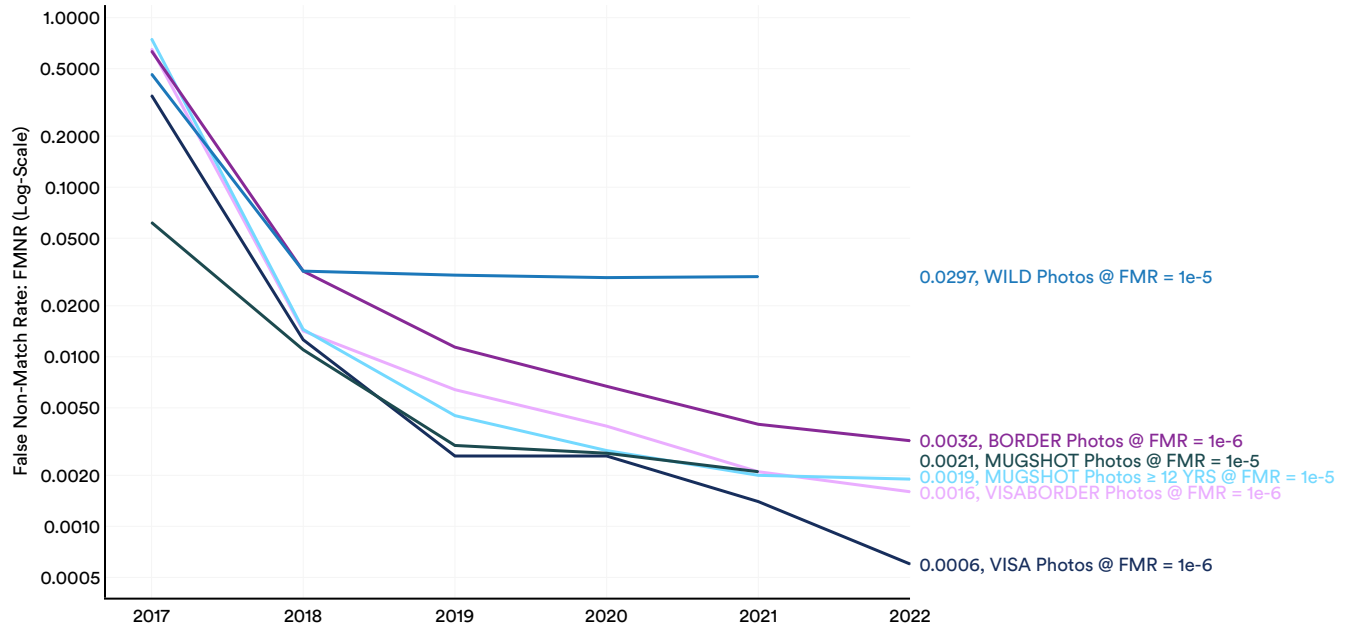


Figure 2.2.4

National Institute of Standards and Technology Face Recognition Vendor Test (FRVT)

Progress on facial recognition can be tracked through the [National Institute of Standards and Technology's Face Recognition Vendor Test](#). This test tracks how well different facial recognition algorithms perform on various homeland security tasks, such as identification of child trafficking victims and cross-verification of visa images, among

others. Facial detection capacity is measured by the false non-match rate (FNMR), otherwise known as error rate, which is the rate at which a model fails to match the face in an image to that of a person.

As of 2022, the top-performing models on all of the FRVT datasets, with the exception of WILD Photos, each posted an error rate below 1%, and as low as a 0.06% error rate on the VISA Photos dataset.

Deepfake Detection

The ability of AI systems to create synthetic images that are sometimes indistinguishable from real ones has led to the creation of deepfakes, images or videos that appear to be real but are actually fake. In the last year, there was a widely circulated deepfake video of Ukrainian president Volodymyr Zelenskyy surrendering (Figure 2.2.5).

Celeb-DF

Celeb-DF is presently one of the most challenging deepfake detection benchmarks. This dataset is composed of 590 original celebrity YouTube videos that have been manipulated into thousands of deepfakes. This year's top deepfake detection

Real-Life Deepfake: President Zelenskyy Calling for the Surrender of Ukrainian Soldiers

Source: [NPR, 2022](#)

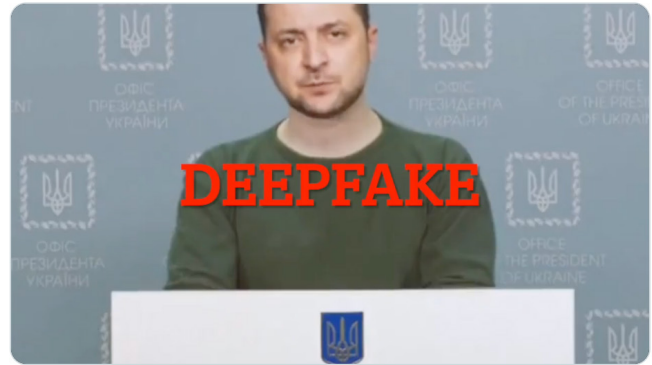


Figure 2.2.5

algorithm on Celeb-DF came from researchers at Deakin University in Australia. Their JDfD model posted an AUC score of 78 (Figure 2.2.6).

Celeb-DF: Area Under Curve Score (AUC)

Source: [arXiv, 2022](#) | Chart: 2023 AI Index Report

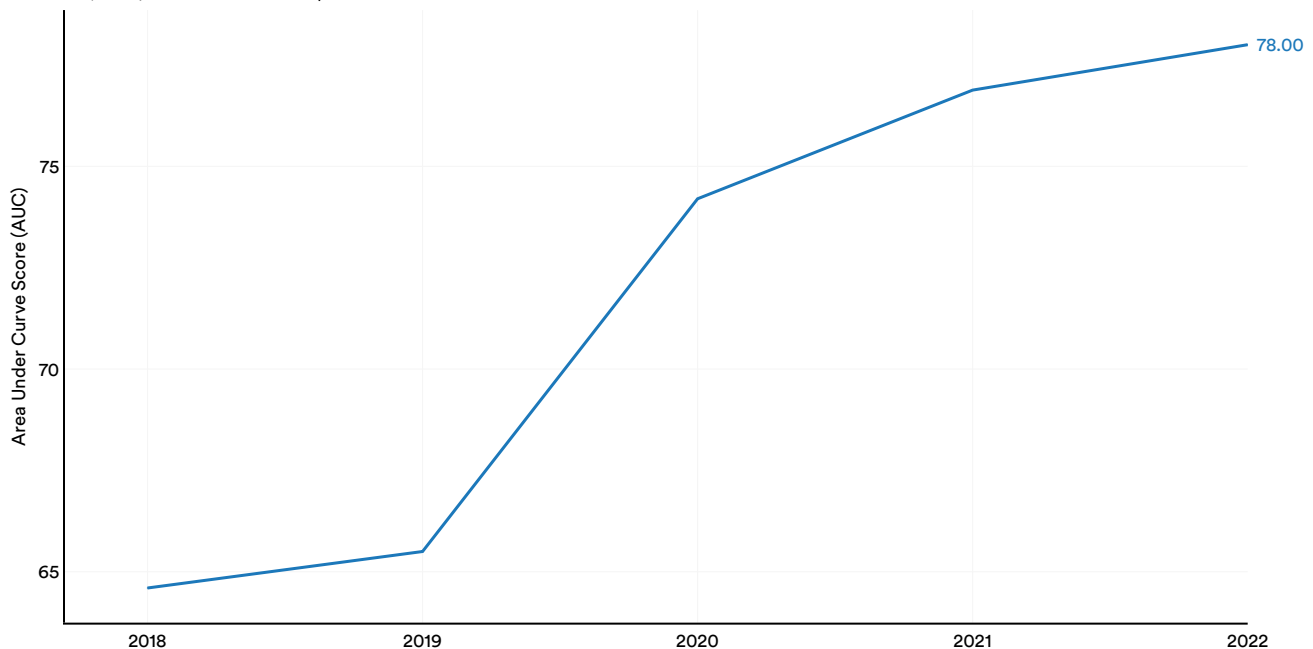


Figure 2.2.6

Human Pose Estimation

Human pose estimation is the task of estimating the position of the human body from images (Figure 2.2.7).

MPII

MPII is a dataset of over 25,000 annotated images which contains annotations of more than 40,000 people doing 410 human activities. On MPII, this year's top model, ViTPose, correctly estimated 94.3% of keypoints (human joints), which represented a small 0.2 percentage point increase from the previous state-of-the-art result posted in 2020 (Figure 2.2.8).

A Demonstration of Human Pose Estimation

Source: [Cong et al., 2022](#)

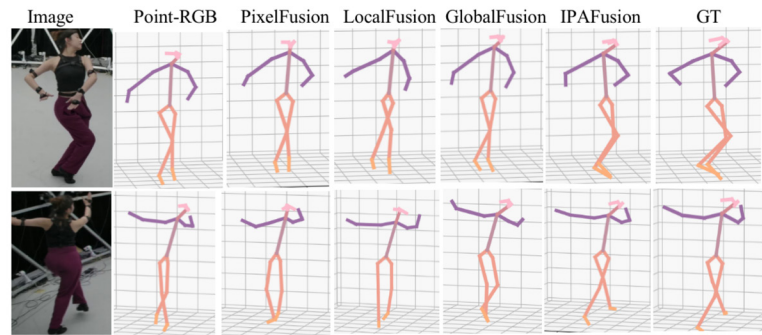


Figure 2.2.7

MPII: Percentage of Correct Keypoints (PCK)

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report

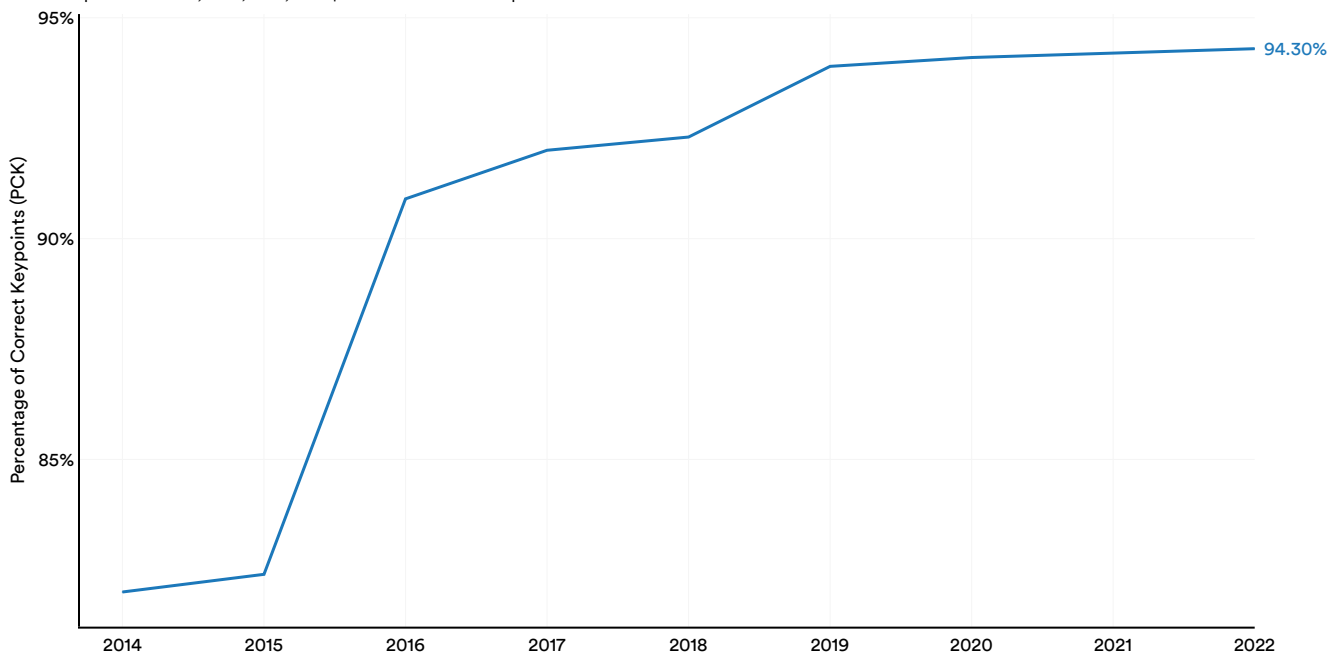


Figure 2.2.8

Semantic Segmentation

Semantic segmentation involves assigning individual image pixels to specific categories (for example, human, bicycle, or street) (Figure 2.2.9).

Cityscapes Challenge, Pixel-Level Semantic Labeling Task

The Cityscapes dataset is used to test the semantic segmentation capabilities of AI. This dataset contains 25,000 annotated images of diverse urban environments. The Cityscapes dataset enables a variety of different segmentation tasks. One of the most popular is the pixel-level task. Performance on semantic segmentation is measured by mean intersection-over-union (mIoU), which represents the degree to which the image segments predicted by the model overlap with the image's actual segments. The

A Demonstration of Semantic Segmentation

Source: Cityscapes Dataset, 2022



Figure 2.2.9

greater the mIoU, the better a system has performed.

Performance on Cityscapes has increased by 23.4 percentage points since the competition launched in 2014; however, it has plateaued in the last few years (Figure 2.2.10).

Cityscapes Challenge, Pixel-Level Semantic Labeling Task: Mean Intersection-Over-Union (mIoU)

Source: Cityscapes Challenge, 2022 | Chart: 2023 AI Index Report

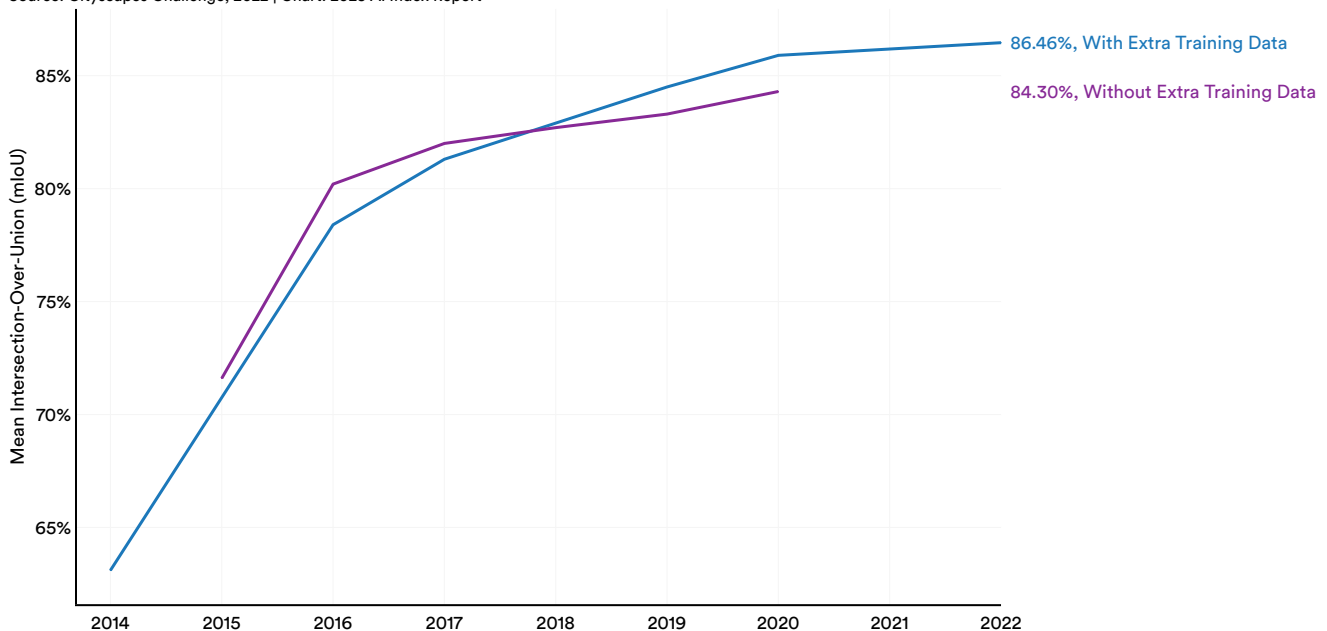


Figure 2.2.10

Medical Image Segmentation

In medical image segmentation, AI systems segment objects such as lesions or organs in medical images (Figure 2.2.11).

Kvasir-SEG

Kvasir-SEG is a dataset for medical image segmentation that contains 1,000 high-quality images of gastrointestinal polyps that were manually identified by medical professionals. Progress on Kvasir-SEG is measured in mean Dice, which represents the degree to which the polyp segments identified by AI systems overlap with the actual polyp segments.¹

A Demonstration of Medical Imaging Segmentation

Source: Jha et al., 2019

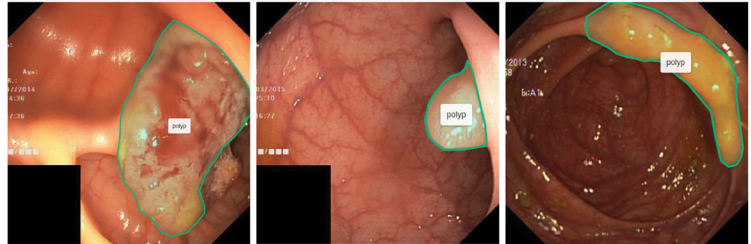


Figure 2.2.11

This year's top-performing model on Kvasir-SEG, SEP, was created by a Chinese researcher and posted a mean Dice of 94.1% (Figure 2.2.12).

Kvasir-SEG: Mean Dice

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report

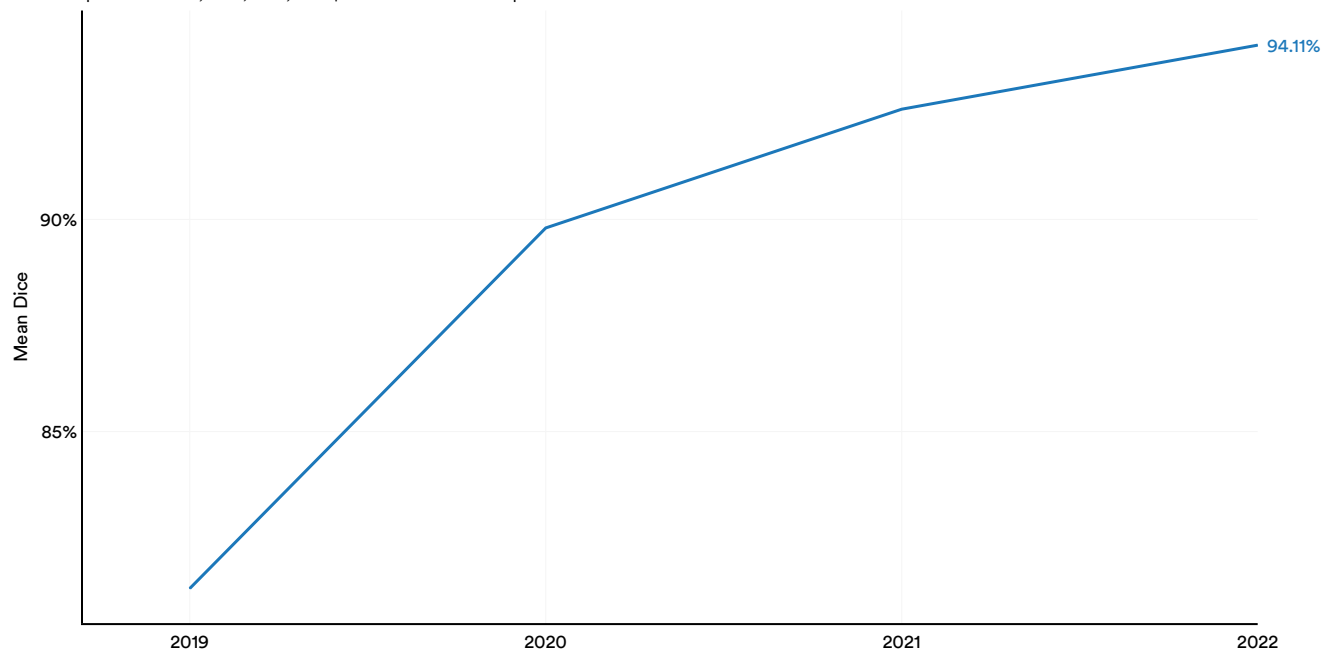


Figure 2.2.12

¹ Mean Dice and mIoU are in principle quite similar. [This StackExchange post](#) outlines the differences in more detail.

Object Detection

The challenge of identifying and localizing objects within an image or video is known as object detection (Figure 2.2.13).

Common Objects in Context (COCO)

Microsoft's Common Objects in Context (COCO) object detection dataset has over 80 object categories in 328,000 images. Several accuracy metrics are used to measure progress on COCO. This section considers mean average precision (mAP50).

Since 2015, state-of-the-art detectors have improved by 26 percentage points. The top model in 2022, EVA, was the result of a Chinese academic research collaboration.

A Demonstration of Object Detection

Source: [Rizzoli, 2023](#)

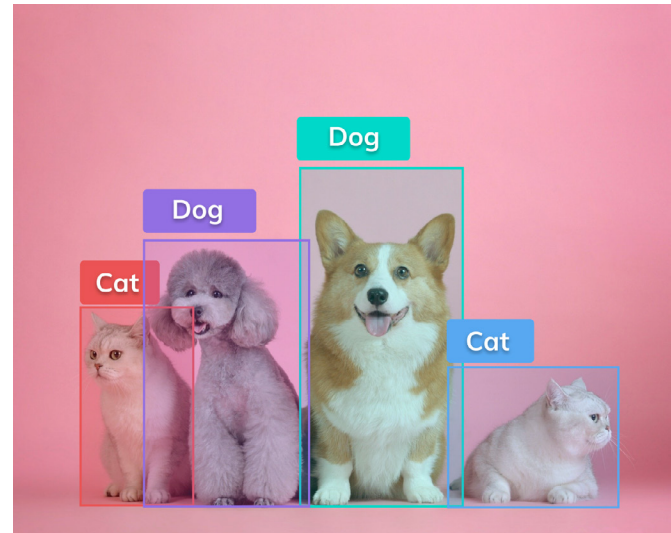


Figure 2.2.13

COCO: Mean Average Precision (mAP50)

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report

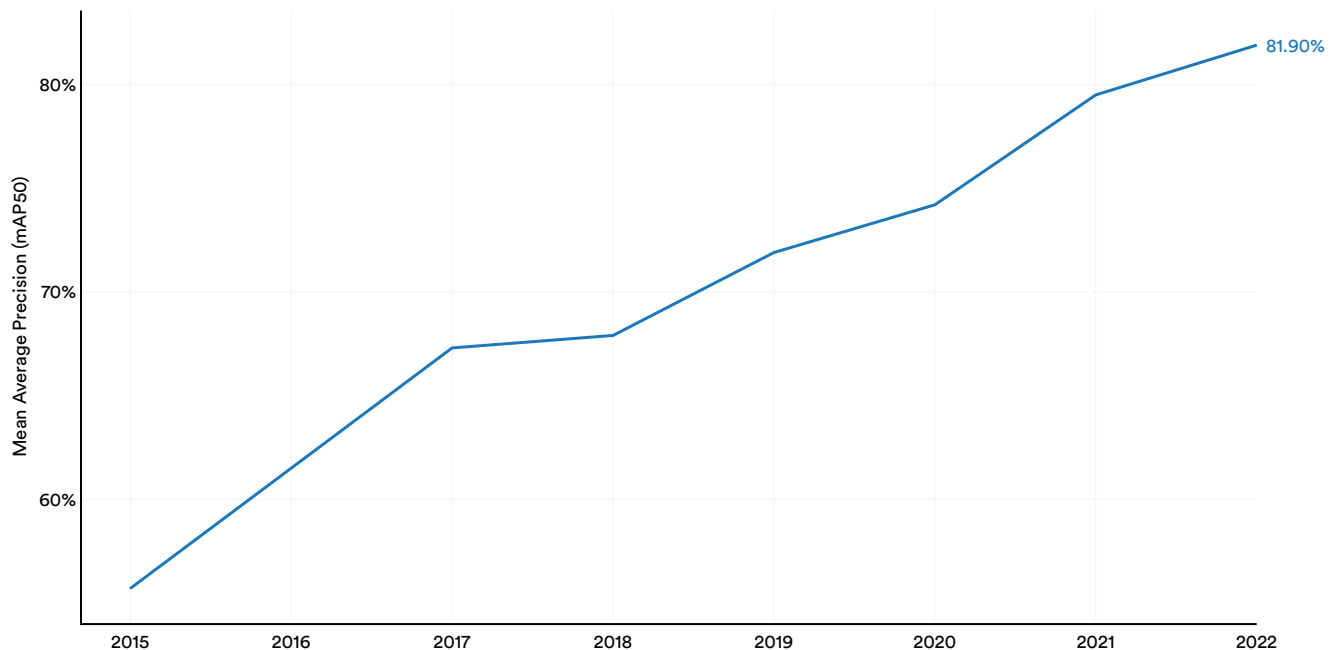


Figure 2.2.14

Image Generation

Image generation is the task of generating images that are indistinguishable from real ones. In the last decade, progress on image generation has tremendously increased, so much so that now it would be difficult for the average person to distinguish a real human face from one synthetically generated by AI (Figure 2.2.15).

CIFAR-10 and STL-10

CIFAR-10 and STL-10 are two popular benchmarks for tracking progress on image generation. CIFAR-10 comprises 60,000 color images across 10 different object classes; STL-10 is inspired by CIFAR-10, with some modifications, including fewer labeled training examples and more unlabeled examples. Progress on image generation in both benchmarks is measured by the Fréchet Inception Distance (FID) score, which reflects the degree to which a synthetically generated

Which Face Is Real?

Source: [Which Face Is Real?, 2022](#)



Figure 2.2.15

set of images is similar to the real images on which it was trained.

This year saw state-of-the-art results on both CIFAR-10 and STL-10 benchmarks (Figure 2.2.16). The top model on CIFAR-10, EDM-G++, came from Korean researchers at KAIST. The top model on STL-10 was Diffusion-GAN, a collaboration between researchers at the University of Texas at Austin and Microsoft.

CIFAR-10 and STL-10: Fréchet Inception Distance (FID) Score

Source: [Papers With Code, 2022](#); [arXiv, 2022](#) | Chart: 2023 AI Index Report

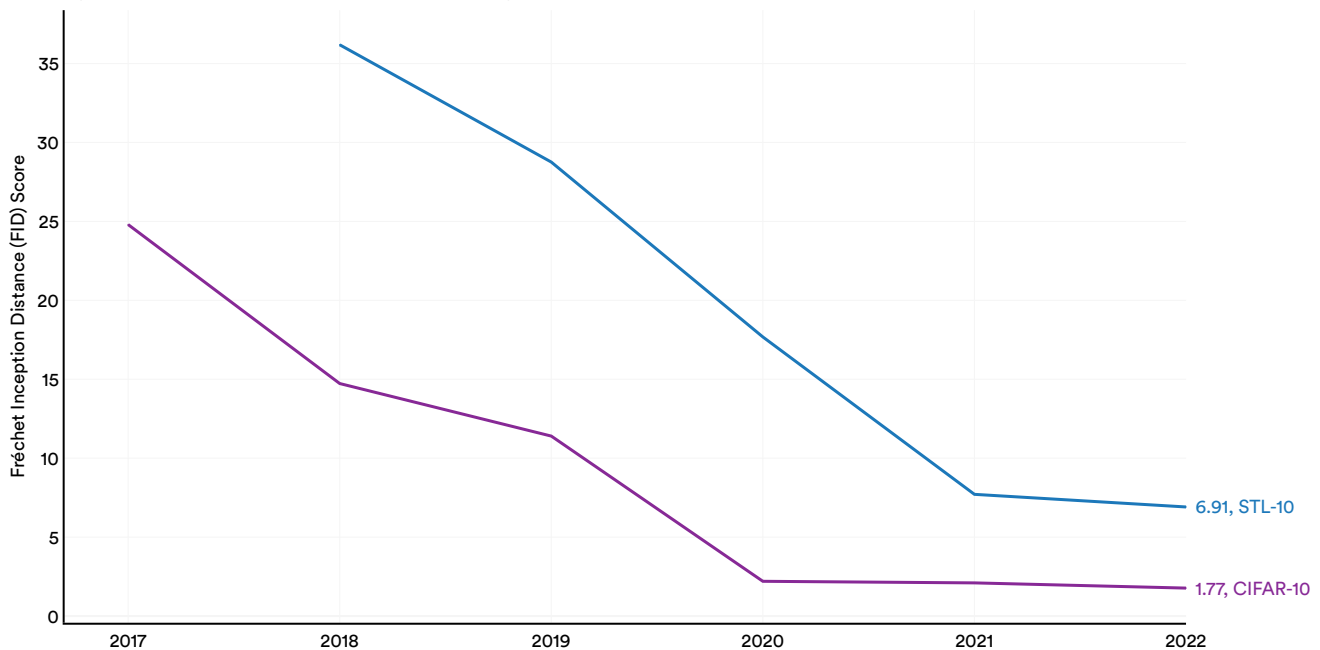


Figure 2.2.16

Narrative Highlight:

A Closer Look at Progress in Image Generation

Figure 2.2.17 tracks the progress of facial image generation over time, with the final image being generated by Diffusion-GAN, the model that posted the 2022 state-of-the-art score on STL-10.

GAN Progress on Face Generation

Source: [Goodfellow et al., 2014](#); [Radford et al., 2016](#); [Liu and Tuzel, 2016](#); [Karras et al., 2018](#); [Karras et al., 2019](#); [Goodfellow, 2019](#); [Karras et al., 2020](#); [Vahdat et al., 2021](#); [Wang et al., 2022](#).

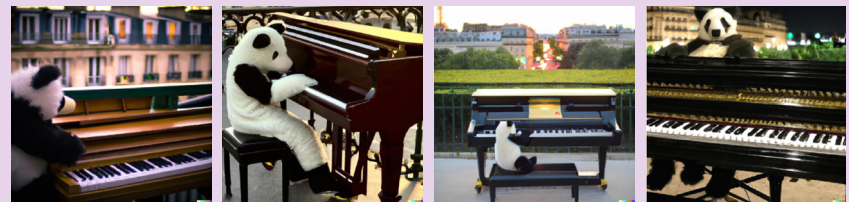


Figure 2.2.17

In the last year, text-to-image generation broke into the public consciousness with the release of models such as OpenAI’s [DALL-E 2](#), Stability AI’s [Stable Diffusion](#), Midjourney’s [Midjourney](#), Meta’s [Make-A-Scene](#), and Google’s [Imagen](#). With these systems, users can generate images based on a text prompt. Figure 2.2.18 juxtaposes the images generated by DALL-E 2, Stable Diffusion, and Midjourney, three publicly accessible AI text-to-image systems, for the same prompt: “a panda playing a piano on a warm evening in Paris.”

Images Generated by DALL-E 2, Stable Diffusion and Midjourney

Source: AI Index, 2022



a. DALL-E 2



b. Stable Diffusion



c. Midjourney

Figure 2.2.18

Narrative Highlight:

A Closer Look at Progress in Image Generation (cont'd)

Of all the recently released text-to-image generators, Google's Imagen performs best on the COCO benchmark (Figure 2.2.19)². This year, the [Google researchers](#) who created Imagen also released a more difficult text-to-image benchmark, DrawBench, designed to challenge increasingly capable text-to-image models.

Notable Text-to-Image Models on MS-COCO 256 × 256 FID-30K: Fréchet Inception Distance (FID) Score

Source: Saharia et al., 2022 | Chart: 2023 AI Index Report

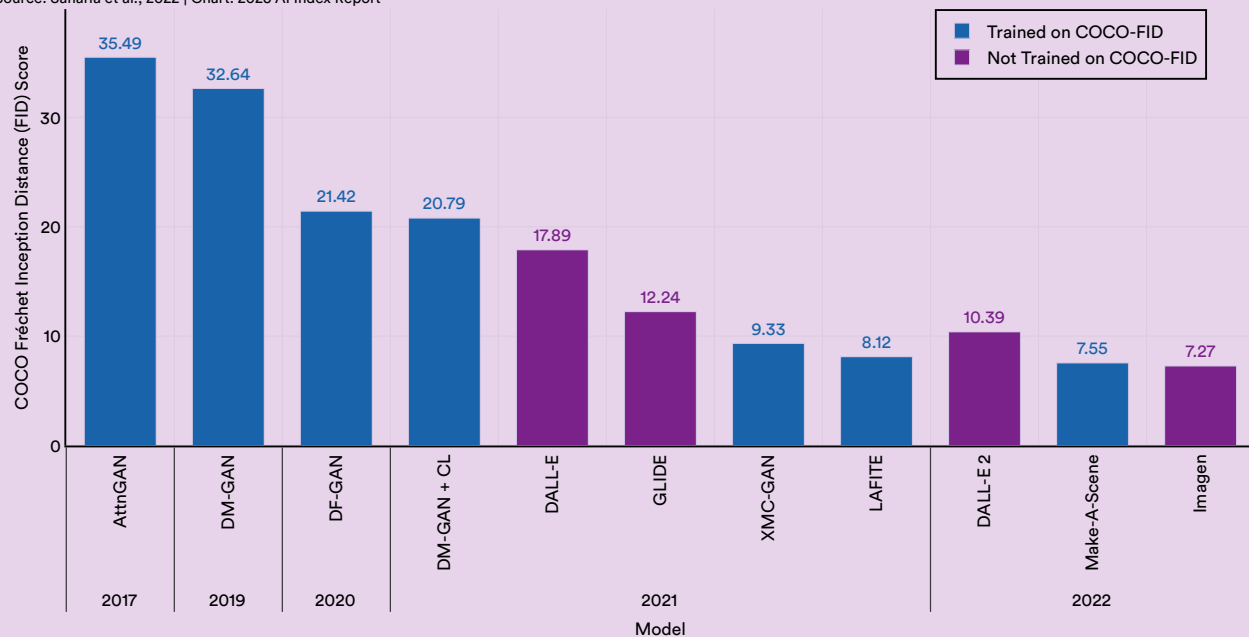


Figure 2.2.19

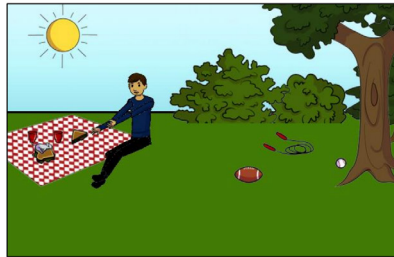
² The COCO benchmark, first launched in 2014, includes 328,000 images with 2.5 million labeled instances. Although it is typically used for object detection tasks, researchers have also deployed it for image generation.

Visual Reasoning

Visual reasoning tests how well AI systems can reason across both textual and visual data, as in the examples of Figure 2.2.20.



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

A Collection of Visual Reasoning Tasks

Source: Agrawal et al., 2016
Figure 2.2.20

Visual Question Answering (VQA) Challenge

The Visual Question Answering Challenge tests AI systems with open-ended textual questions about images. Successfully answering the questions requires that AI systems possess vision, language, and commonsense reasoning capabilities. This section

reports progress on the VQA V2 dataset.

This year the top-performing model on VQA V2 was PaLI, a multimodal model produced by Google researchers (Figure 2.2.21).

Visual Question Answering (VQA) V2 Test-Dev: Accuracy

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report

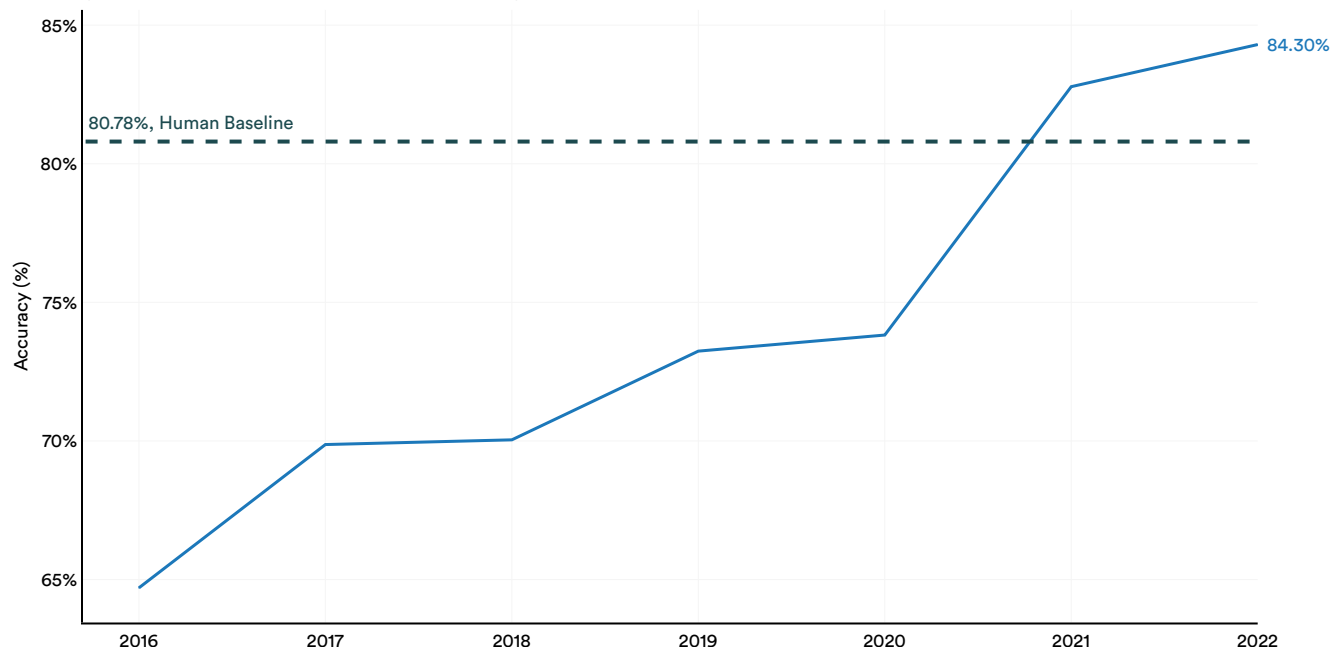


Figure 2.2.21



Narrative Highlight:

The Rise of Capable Multimodal Reasoning Systems

Traditionally AI has been strong in narrow tasks, but it has been unable to easily generalize across multiple domains. For instance, many image classifiers are adept at classifying images but are incapable of understanding written text.

However, recent technical progress in AI has begun to challenge this notion. In 2022, several

models were introduced, for example [BEiT-3](#) from Microsoft and [PaLI](#) from Google, that posted state-of-the-art results across a variety of both vision and language benchmarks. For example, at the time of publication of the BEiT-3 paper, BEiT-3 posted state-of-the-art results for four different vision skills and five different vision-language skills (Figure 2.2.22).

BEiT-3 Vs. Previous State-of-the-Art Models

Source: Wang et al., 2022 | Table: 2023 AI Index Report

Category	Task	Dataset	Metric	Previous SOTA	Model of Previous SOTA	BEiT-3	Scale of Improvement
Vision	Semantic Segmentation	ADE20K	mIoU	61.40	FD-SwimV2	62.80	2.28%
Vision	Object Detection	COCO	AP	63.30	DINO	63.70	0.63%
Vision	Instance Segmentation	COCO	AP	54.70	Mask DINO	54.80	0.18%
Vision	Image Classification	ImageNet	Top-1 Accuracy	89.00	FD-CLIP	89.60	0.67%
Vision-Language	Visual Reasoning	NLVR	Accuracy	87.00	CoCA	92.60	6.44%
Vision-Language	Visual QA	VQAv2	VQA Accuracy	82.30	CoCA	84.00	2.07%
Vision-Language	Image Captioning	COCO	CIDEr	145.30	OFA	147.60	1.58%
Vision-Language	Finetuned Retrieval	COCO Flickr30K	R@1	72.50	Florence	76.00	4.83%
Vision-Language	Zero-Shot Retrieval	Flickr30K	R@1	86.50	CoCA	88.20	1.97%

Figure 2.2.22

Narrative Highlight:

The Rise of Capable Multimodal Reasoning Systems (cont'd)

Figure 2.2.23 shows some of the different vision-language tasks challenging multimodal systems like PaLI and BEiT-3.

A Collection of Vision-Language Tasks

Source: [Chen et al., 2022](#)



Input: Generate the alt_text in EN
Output: A cellar filled with barrels of wine



Input: Generate the alt_text in EN
Output: a clock on a building that says 'lvania' on it



Input: Generate the alt_text in EN
Output: Two helicopters are flying in the sky and one has a yellow stripe on the tail



Input: Generate the alt_text in FR
Output: Un arbre debout dans un champ avec un ciel violet



Input: Generate the alt_text in TH
Output: ม้าลายยืนในทุ่งหญ้า



Input: Generate the alt_text in ZH
Output: 一辆电动汽车停在充电桩上。



Input: Answer in EN: what time is it according to this radio
Output: 1254



Input: Answer in EN: what website is on the wall in back
Output: arsenaldirect.com



Input: Answer in EN: what is the brand of this watch
Output: seiko

Figure 2.2.23

Visual Commonsense Reasoning (VCR)

The Visual Commonsense Reasoning challenge, first launched in 2019, is a relatively new benchmark in which AI systems must answer questions presented from images, as in VQA, but also select the reasoning behind their answer choices. Figure 2.2.24 shows an

example of a question posed in VCR. Performance on VCR is tracked in the Q->AR score, which combines the ability of machines to select the right answer for the question (Q->A) and the ability to select the correct rationale behind the answer (Q->R).

A Sample Question from the Visual Commonsense Reasoning (VCR) Challenge

Source: Zellers et al., 2018



How did [person2] get the money that's in front of her?

- a) [person2] is selling things on the street.
 - b) [person2] earned this money playing music.
 - c) She may work jobs for the mafia.
 - d) She won money playing poker.
- I chose b) because...
- a) She is playing guitar for money.
 - b) [person2] is a professional musician in an orchestra.
 - c) [person2] and [person1] are both holding instruments, and were probably busking for that money.
 - d) [person1] is putting money in [person2]'s tip jar, while she plays music.

Figure 2.2.24

VCR is one of the few visual benchmarks considered in this report on which AI systems have yet to surpass human performance, as shown in Figure 2.2.25.

Visual Commonsense Reasoning (VCR) Task: Q->AR Score

Source: VCR Leaderboard, 2022 | Chart: 2023 AI Index Report

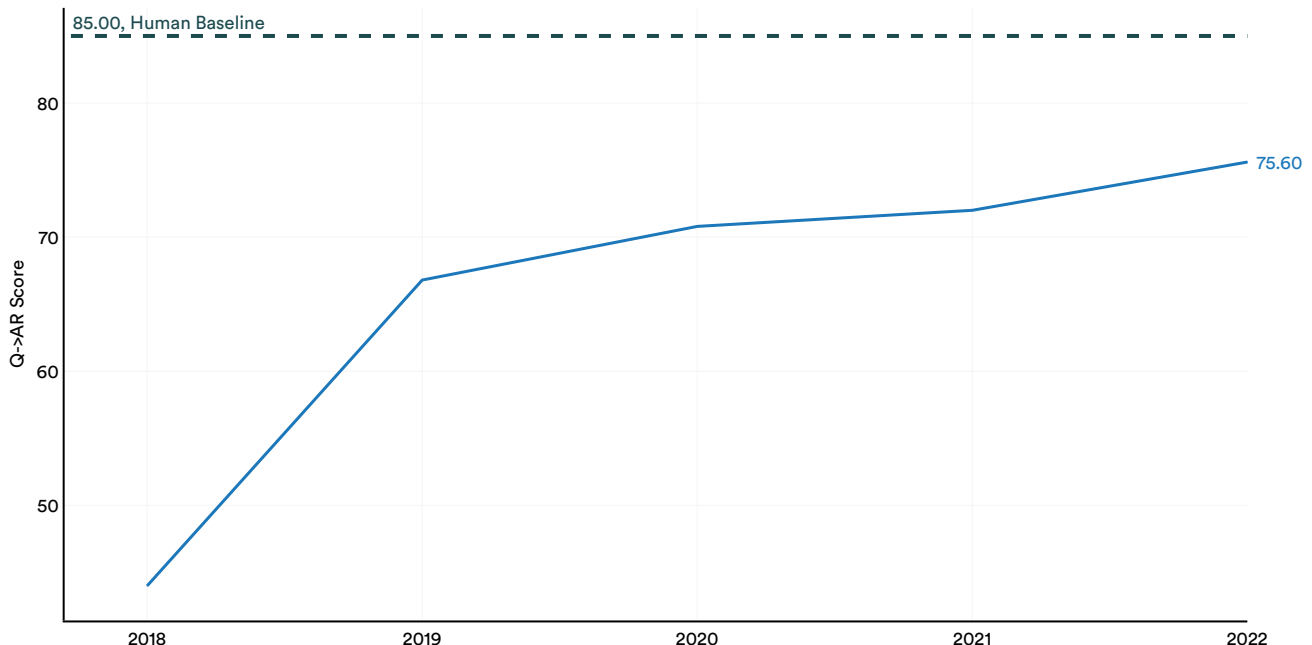


Figure 2.2.25

Video analysis concerns reasoning or task operation across videos, rather than single images.

2.3 Computer Vision—Video

Activity Recognition

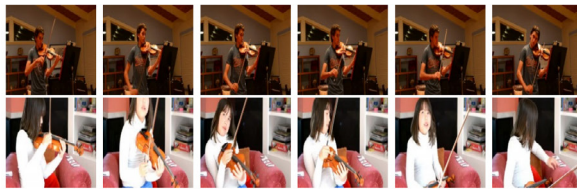
Activity recognition is the categorization of activities that occur in videos. Certain activities, such as sitting, sleeping, or walking, are easier for AI systems to categorize than others which involve multiple steps—for example, preparing dinner.

Kinetics-400, Kinetics-600, Kinetics-700

Kinetics-400, Kinetics-600, and Kinetics-700 are a series of datasets for benchmarking video activity recognition. Each dataset includes 650,000 large-scale, high-quality video clips from YouTube that display a wide range of human activities, and each asks AI systems to classify an action from a possible set of 400, 600, and 700 categories, respectively (Figure 2.3.1).

Example Classes From the Kinetics Dataset

Source: Kay et al., 2017



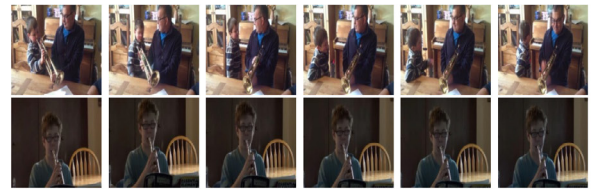
(i) playing violin



(k) braiding hair



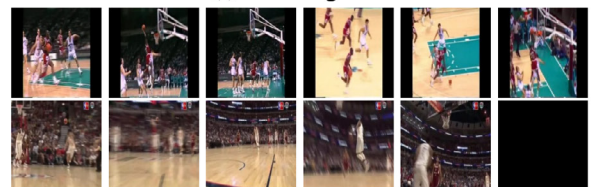
(m) dribbling basketball



(j) playing trumpet



(l) brushing hair



(n) dunking basketball

Figure 2.3.1

As of 2022, there is a 7.8 percentage point gap in performance between the top system on Kinetics-600 and Kinetics-700, which suggests the 700 series dataset is still a meaningful challenge for video computer vision researchers (Figure 2.3.2).

Kinetics-400, Kinetics-600, Kinetics-700: Top-1 Accuracy

Source: Papers With Code, 2021; arXiv, 2022 | Chart: 2023 AI Index Report

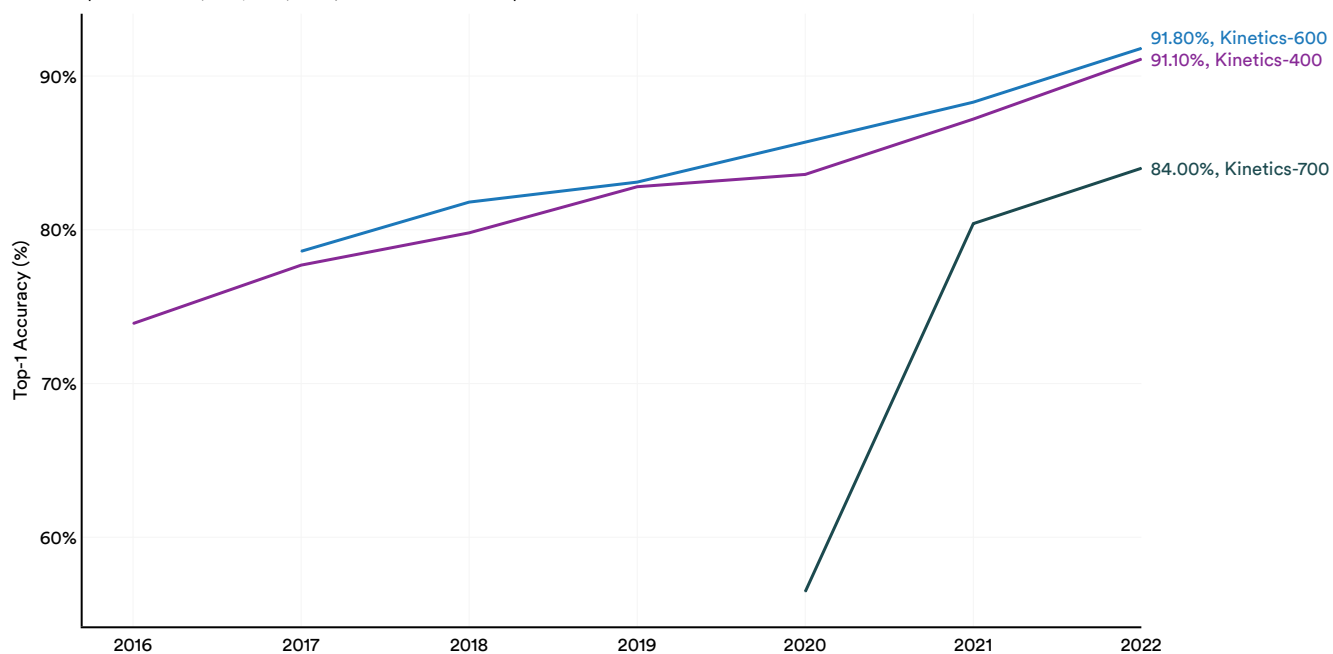


Figure 2.3.2

Narrative Highlight:

A Closer Look at the Progress of Video Generation

Multiple high quality text-to-video models, AI systems that can generate video clips from prompted text, were released in 2022³. In May, researchers from Tsinghua University and the Beijing Academy of Artificial Intelligence released [CogVideo](#), a model that posted the then-highest inception score on the UCF-101 benchmark for text-to-video generation (Figure 2.3.3).

In September 2022, CogVideo's top score was significantly surpassed by Meta's [Make-A-Video](#) model (Figure 2.3.3). Make-A-Video performed 63.6% better on UCF-101 than CogVideo. And, in October 2022, Google released a text-to-video system called [Phenaki](#); however, this model was not benchmarked on UCF-101.

Notable Text-to-Video Models on UCF-101: Inception Score (IS)

Source: Hong et al., 2022; Singer et al., 2022 | Chart: 2023 AI Index Report

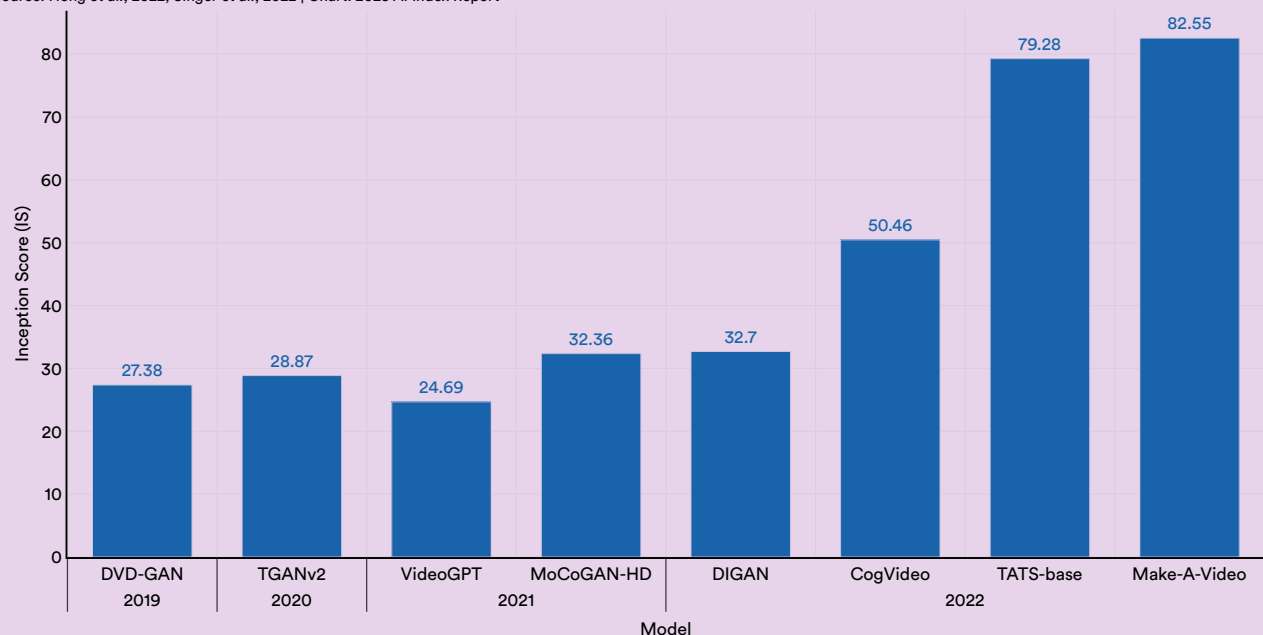


Figure 2.2.3

³ Although these models are impressive, it is worth noting that they are thus far only capable of generating videos of a few seconds' duration.

Natural language processing (NLP) is the ability of computer systems to understand text. The last few years have seen the release of increasingly capable “large language models,” AI systems like PaLM, GPT-3, and GLM-130B, that are trained on massive amounts of data and adaptable to a wide range of downstream tasks.

In this section, progress in NLP is tracked across the following skill categories: (1) English language understanding, (2) text summarization, (3) natural language inference, (4) sentiment analysis, (5) multitask language understanding, and (6) machine translation.

2.4 Language

English Language Understanding

English language understanding challenges AI systems to understand the English language in various ways: reading comprehension, yes/no reading comprehension, commonsense reading comprehension, and logical reasoning.

SuperGLUE

SuperGLUE is a comprehensive English language understanding benchmark that tracks the progress of AI models on eight different linguistic tasks.

A selection of these tasks is highlighted in Figure 2.4.1. Their performance is then aggregated into a single metric.

A Set of SuperGLUE Tasks⁴

Source: Wang et al., 2019

ReCoRD	<p>Paragraph: (<i>CNN</i>) <i>Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electoral Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood</i></p> <p>Query For one, they can truthfully say, “Don’t blame me, I didn’t vote for them,” when discussing the <placeholder> presidency Correct Entities: US</p>
RTE	<p>Text: <i>Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.</i></p> <p>Hypothesis: <i>Christopher Reeve had an accident.</i> Entailment: False</p>
WiC	<p>Context 1: <i>Room and board.</i> Context 2: <i>He nailed boards across the windows.</i></p> <p>Sense match: False</p>
WSC	<p>Text: <i>Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.</i> Coreference: False</p>

Figure 2.4.1

⁴ For the sake of brevity, this figure only displays four of the eight tasks.



This year's top model on SuperGLUE, Vega, registered a new state-of-the-art score of 91.3, which is 1.5 percentage points higher than the human baseline. Performance on SuperGLUE is continuing to saturate.

SuperGLUE: Score

Source: SuperGLUE Leaderboard, 2022 | Chart: 2023 AI Index Report

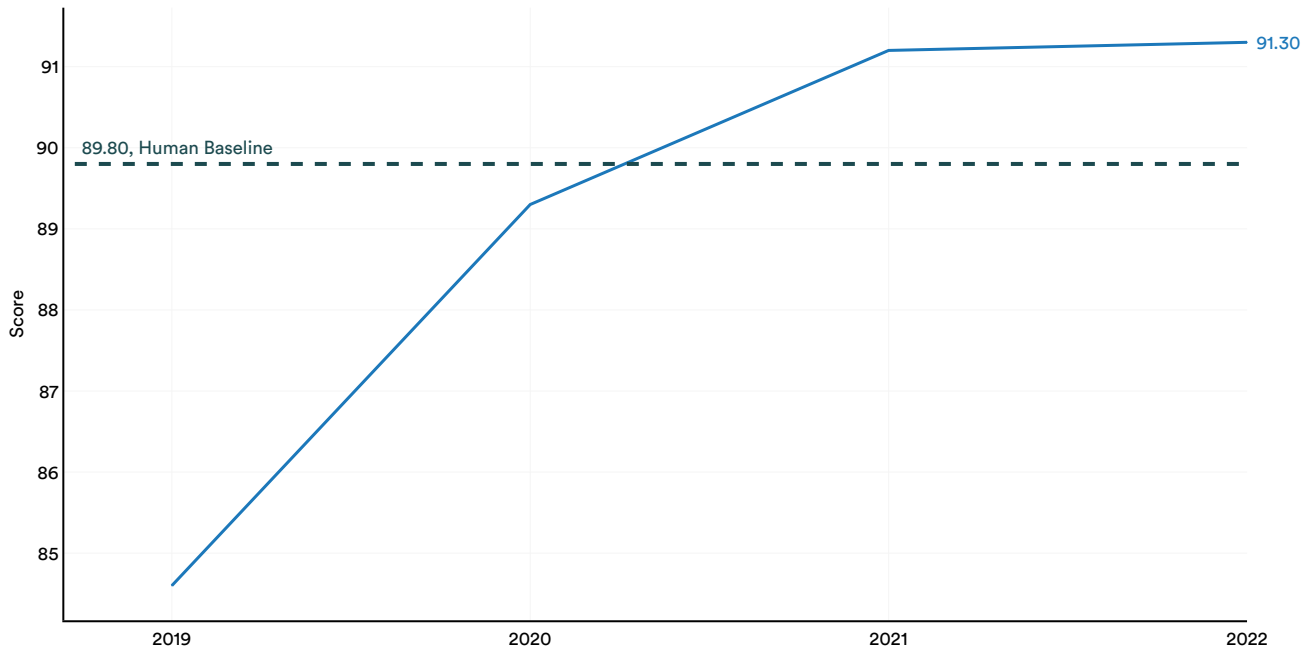


Figure 2.4.2

Reading Comprehension Dataset Requiring Logical Reasoning (ReClor)

In response to the saturation of traditional reading comprehension benchmarks, researchers from the National University of Singapore launched ReClor in 2020. ReClor, or Reading Comprehension Dataset Requiring Logical Reasoning, is a dataset of logical reasoning questions taken from the LSAT, the entrance exam for law schools in the United States and Canada. A sample question is shown in Figure 2.4.3

A Sample Question from the Reading Comprehension Dataset Requiring Logical Reasoning (ReClor)

Source: Yu et al., 2020

Context: When a certain gland becomes cancerous in humans, it produces high levels of a particular protein. A blood test can determine the level of this protein well before a cancer of the gland could be detected by other means. Some doctors recommend that aggressive anticancer treatment should be begun as early as possible for anyone who is tested and is found to have high levels of the protein.

Question: Which one of the following, if true, most seriously weakens the doctors' recommendation?

- A.** The blood test for the protein has been in use for some time to monitor the condition of patients who have been diagnosed as having cancer of the gland.
- B.** Before the blood test became available, about one-third of all cases of cancer of the gland were detected in early stages.
- C.** So far, no patients whose protein levels were found to be normal have subsequently developed cancer of the gland.
- D.** Enlargement of the gland, a common condition infrequently associated with cancer, results in high levels of the protein.

Figure 2.4.3



Figure 2.4.4 examines progress on ReClor. The top 2022 result of 80.6% represented an 18 percentage point improvement from 2020, the year the benchmark was released.

Reading Comprehension Dataset Requiring Logical Reasoning (ReClor): Accuracy

Source: ReClor Leaderboard, 2022; Papers With Code, 2022 | Chart: 2023 AI Index Report

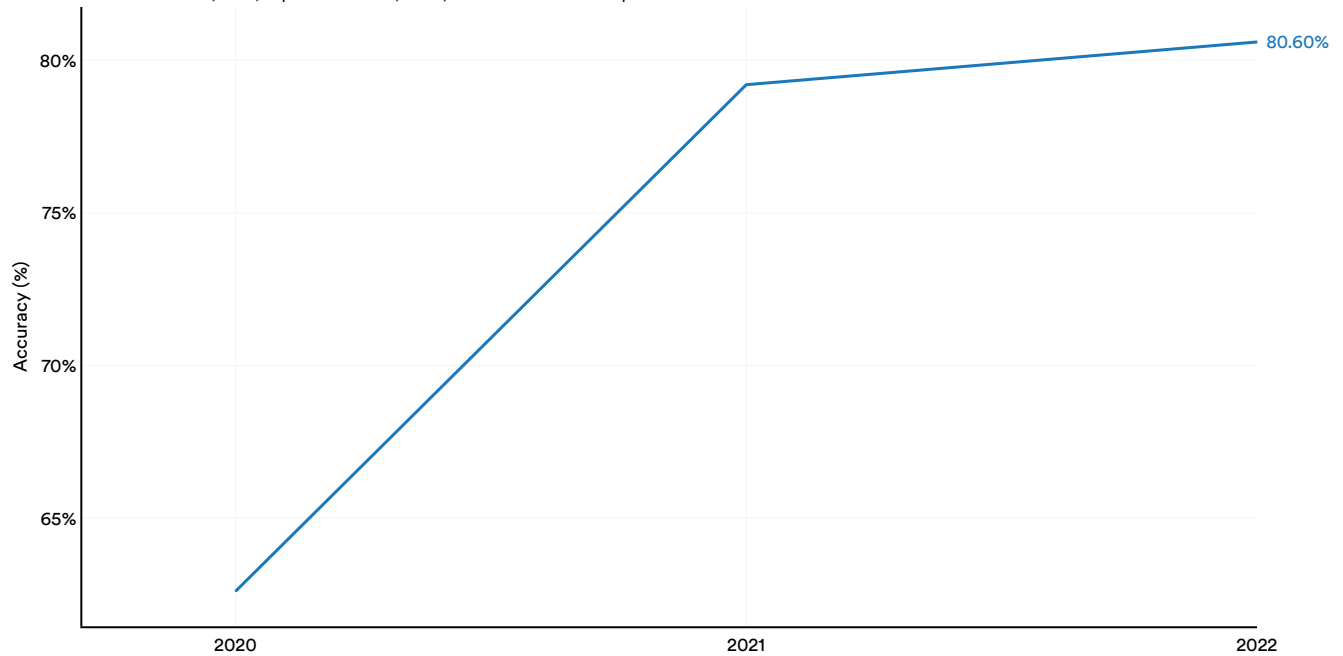


Figure 2.4.4



Narrative Highlight:

Just How Much Better Have Language Models Become?

The AI Index tested how three large language models from three different years, GPT-2 (2019), GPT-3 (2020), and ChatGPT (2022), handle the same prompt: “Explain to me the major accomplishments of Theodore Roosevelt’s presidency.” More recent models are able to answer this question more effectively, both in terms of factual accuracy and quality of writing.

GPT-2 (February 2019)⁵

(The most notable is the establishment of a national telephone and mail exchange company.)

But the only things that will change with a president who knows how to operate a business is his understanding of the American system of government. The public knows all about our problems, yet at least half of the people who need it won’t know anything about ours. We can no longer have this. Our system of government is the only one in which you really need a public relations firm to work hard.

The system of governmental control, while working, is not always effective in ensuring that the very things that our economy needs, do not always be at the top of everybody’s list. So I wanted to know how successful the system you employ for managing the public’s needs is in helping the American economy grow, how efficient the service our economy provides is being made, and how it can better serve the needs of a growing number of Americans.

I wanted to know.

This was before I ever hired my own lawyer to work on the case. This was before I ever hired my own law firm to handle government investigations. For my first major job, at least, I had to convince two top aides to get paid \$90,000 for the role, and before I even started on the case, I had to convince one top staffer to get paid even \$100,000 to have an interview with my lead prosecutor...⁶

Commentary: Most of this answer is gibberish. GPT-2 is able to respond with sentences that are largely grammatical, but fails to actually answer the question in any meaningful way.

GPT-3 (June 2020)⁷

Theodore Roosevelt’s presidency was marked by a number of accomplishments. He is credited with leading the United States out of the Great Depression and into World War II. He also spearheaded the creation of the National Park Service and the National Wildlife Federation. Roosevelt also played a major role in the development of the Panama Canal.

Commentary: This answer is an improvement over that of GPT-2 in that it actually somewhat responds to the question, but still shows significant flaws. GPT-3 is correct in asserting that Theodore Roosevelt was instrumentally involved in the development of the Panama Canal. However, he did not lead the United States out of the Great Depression and into World War II—that was done by a different American president, Franklin Delano Roosevelt. It appears that GPT-3 confused some of Franklin Delano Roosevelt’s accomplishments with those of Theodore Roosevelt. Finally, although Theodore Roosevelt did play a significant role in establishing many national parks, he did not in fact create the National Park Service or the National Wildlife Federation. The National Park Service was created in 1916 by Woodrow Wilson, and the National Wildlife Federation in 1936, much later than Theodore Roosevelt’s presidency.⁸

ChatGPT (November 2022)

Theodore Roosevelt, President from 1901-1909, accomplished the following major accomplishments during his presidency:

1. Conservation: Established numerous national parks, monuments, and forests, including the Grand Canyon National Park.
2. Trust-busting: Took on monopolies, such as Standard Oil, through antitrust laws and broke up large corporations.
3. Labor Reforms: Improved working conditions for Americans through advocacy for better hours, wages, and workplace safety.
4. Progressive Reforms: Promoted progressive policies, including regulation of big business and protection of natural resources.
5. National Infrastructure: Improved national infrastructure through construction of the Panama Canal and funding for dam projects.
6. Diplomacy: Negotiated the end of the Russo-Japanese War, resulting in his being awarded the Nobel Peace Prize.
7. Foreign Policy: Established the Roosevelt Corollary, asserting U.S. intervention in Latin America to maintain stability and protect American interests.

Commentary: ChatGPT’s response is the strongest of the three. All but one of the facts about Theodore Roosevelt’s presidency are correct. Roosevelt proclaimed the Grand Canyon Game Preserve in 1906 and designated it a national monument in 1908. However, the Grand Canyon National Park was itself created by Woodrow Wilson.⁹ Otherwise the response is factually accurate. The response is also more thorough and concise than the answers provided by GPT-2 and GPT-3.

⁵ GPT-2 used the 124M parameter model downloaded from [OpenAI’s GitHub page](#).

⁶ The complete answer outputted by GPT-2 is trimmed here for brevity. The full answer is included in the Appendix.

⁷ The specific GPT-3 model that was used was text-curie-001, which has training data up to October 2019.

⁸ The information in this section has been cross-verified with the Encyclopedia Britannica entries on [Theodore Roosevelt](#), [Franklin Delano Roosevelt](#), [Woodrow Wilson](#), and the [National Park Service](#), as well as the history page of the National Wildlife Federation.

⁹ Information on the history of the [Grand Canyon National Park](#) was cross-verified with the Wikipedia entry on the Grand Canyon National Park.

Narrative Highlight:

Planning and Reasoning in Large Language Models

As illustrated above, AI systems have become increasingly strong on a wide range of reasoning tasks. This improvement has led many to claim that emerging AI systems, especially large language models, possess reasoning abilities that are somewhat similar to those possessed by humans.¹⁰ Other authors, however, have argued otherwise.¹¹

In 2022, researchers (Valmeekam et al., 2022) introduced a more challenging planning and reasoning test for large language models that consists of seven assignments: (1) plan generation, (2) cost-optimal planning, (3) reasoning about plan execution, (4) robustness to goal reformulation, (5) ability to reuse plans, (6) replanning, and (7) plan generalization.¹²

The authors then tested notable language models on these tasks in a Blocksworld problem domain, a problem environment where agents are given blocks of different colors and tasked with arranging these blocks in particular orders. The authors demonstrated that these large language models performed fairly ineffectively (Figure 2.4.5). While GPT-3, Instruct-GPT3, and BLOOM demonstrated the ability, in some contexts, to reformulate goals in robust ways, they struggled with other tasks like plan generation, optimal planning, and plan reuse. Compared to humans, the large language models performed much worse, suggesting that while they are capable, they lack human reasoning capabilities.

Select Large Language Models on the Blocksworld Domain: Instances Correct

Source: Valmeekam et al., 2022 | Chart: 2023 AI Index Report

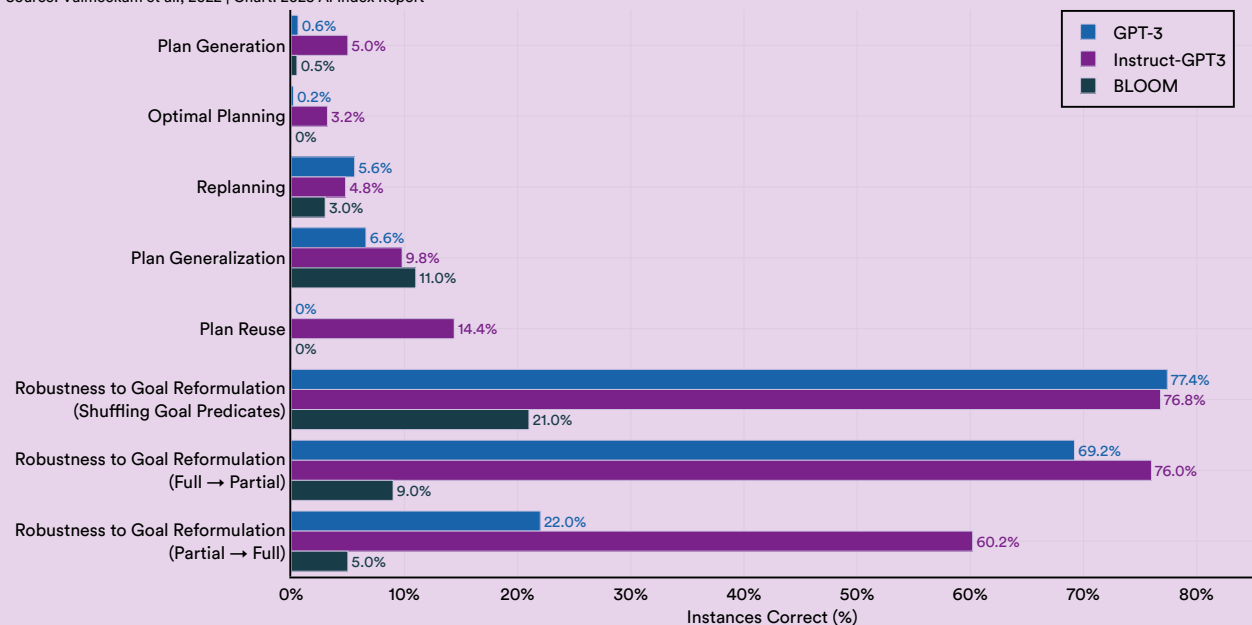


Figure 2.4.5

¹⁰ Some of the papers that claim language models can reason include: [Kojima et al., 2022](#); [Chowdhery et al., 2022](#); [Li et al., 2021](#); [Wei et al., 2022](#).

¹¹ Valmeekam et al., 2022 advances this claim.

¹² A complete description of these tasks can be found in the paper.

Text Summarization

Text summarization tests how well AI systems can synthesize a piece of text while capturing its core content. Text summarization performance is judged on ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which measures the degree to which an AI-produced text summary aligns with a human reference summary.

arXiv and PubMed

arXiv and PubMed are two widely used datasets for benchmarking text summarization. The model that posted the state-of-the-art score in 2022 on both arXiv and PubMed, AdaPool, was developed by a team from Salesforce Research (Figure 2.4.6).

ArXiv and PubMed: ROUGE-1

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report

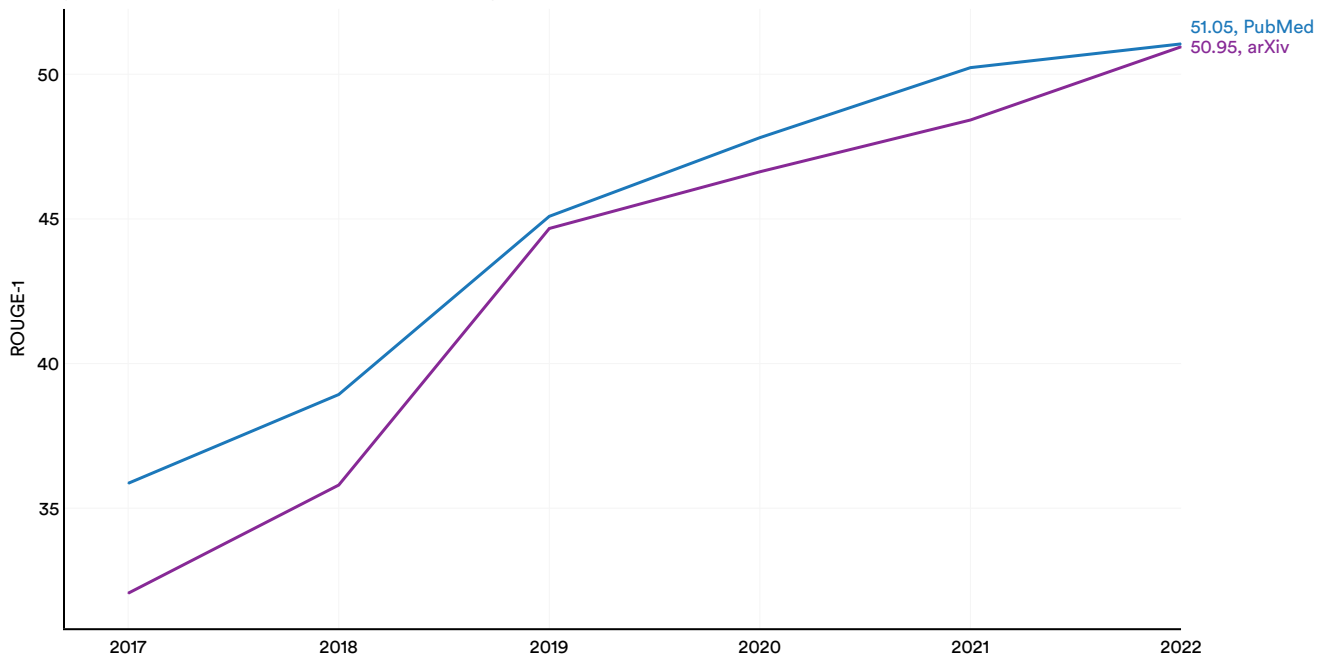


Figure 2.4.6

Natural Language Inference

Also known as textual entailment, natural language inference is the ability of AI systems to determine whether a hypothesis is true, false, or undetermined based on presented premises.

Abductive Natural Language Inference (aNLI)

Abductive natural language inference is a form of natural language inference in which plausible conclusions must be drawn from a set of limited and

uncertain premises. Imagine, for example, that Peter returns to his car after dinner at a restaurant to find the window shattered and his laptop, which he left in the back seat, missing. He might immediately conclude that a thief broke into his car and stole the laptop.

In 2019, the Allen Institute for AI launched [aNLI](#), a comprehensive benchmark for abductive natural language inference that includes 170,000 premise and hypothesis pairs (Figure 2.4.7).

Sample Question From the Abductive Natural Language Inference Benchmark (aNLI)

Source: [Allen Institute for AI, 2021](#)

Obs1: Jenny was addicted to sending text messages.

Obs2: Jenny narrowly avoided a car accident.

Hyp1: Since her friend's texting and driving car accident, Jenny keeps her phone off while driving.

Hyp2: Jenny was looking at her phone while driving so she wasn't paying attention.

Figure 2.4.7

Abductive natural language inference is a challenging task. The human baseline remained unsurpassed until 2022, when an AI system registered a score of 93.7% (Figure 2.4.8).

Abductive Natural Language Inference (aNLI): Accuracy

Source: Allen Institute for AI, 2022 | Chart: 2023 AI Index Report

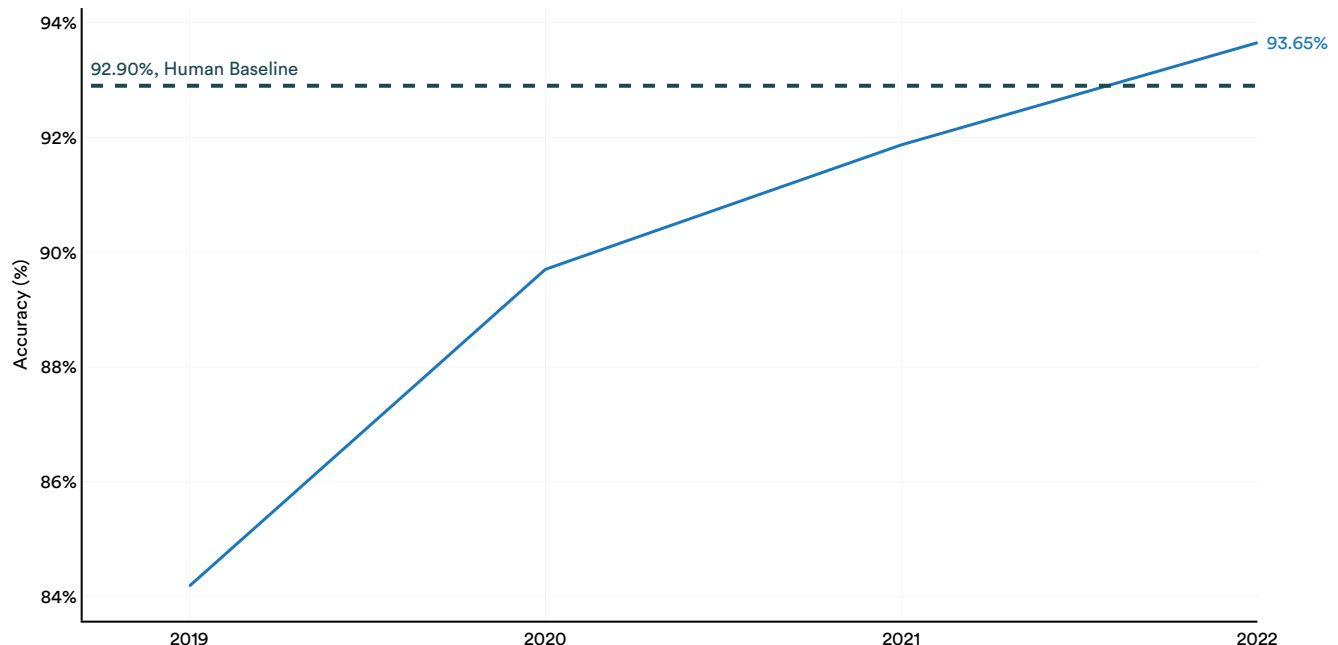


Figure 2.4.8

Sentiment Analysis

Sentiment analysis applies NLP techniques to identify the sentiment of a particular text. It is used by many businesses to better understand customer reviews.

SST-5 Fine-Grained Classification

The Stanford Sentiment Treebank (SST) is a dataset of 11,855 single sentences taken from movie reviews that are then transformed into 215,154 unique phrases whose sentiments have been annotated by human judges (Figure 2.4.9).

A Sample Sentence from SST

Source: Socher et al., 2013

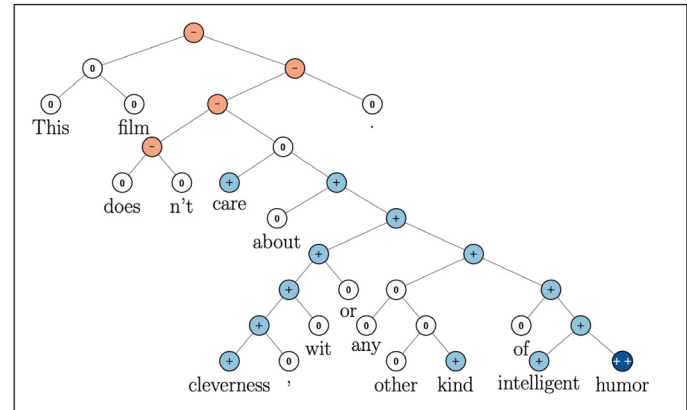


Figure 2.4.9

A new state-of-the-art score of 59.8% was posted on SST-5 fine-grained classification by the Heinsen Routing + RoBERTa Large model (Figure 2.4.10).

SST-5 Fine-Grained: Accuracy

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report

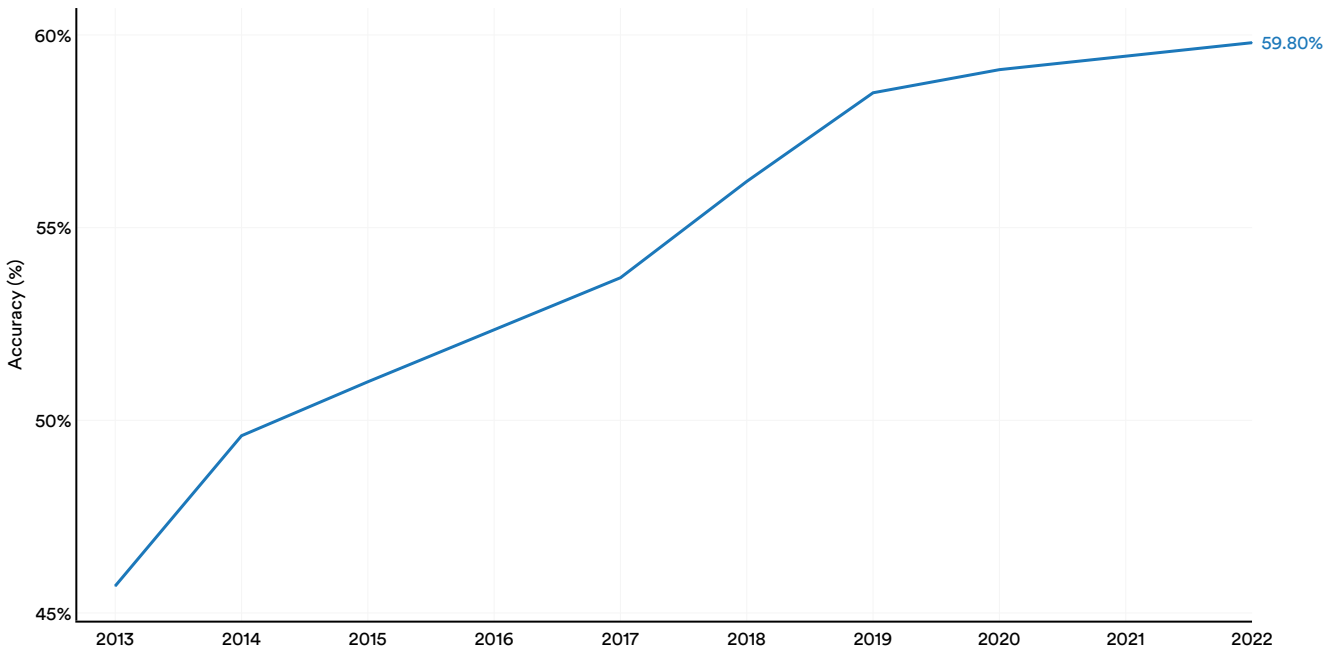


Figure 2.4.10

Multitask Language Understanding

A common criticism of language benchmarks such as GLUE and SuperGLUE is that they do not accurately test how capable language models are at applying the knowledge they learn across different domains.¹³ Multitask language understanding tests the ability of language models to reason across specialized subject domains.

Massive Multitask Language Understanding (MMLU)

Massive Multitask Language Understanding (MMLU) evaluates models in zero-shot or few-shot settings across 57 diverse subjects in the humanities, STEM, and the social sciences (Figure 2.4.11).

Sample Questions From MMLU

Source: Hendrycks et al., 2021

a) Sample Math Questions

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?

(A) 75 (B) 76 (C) 22 (D) 23

Answer: B

Compute $i + i^2 + i^3 + \dots + i^{258} + i^{259}$.

(A) -1 (B) 1 (C) i (D) $-i$

Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?

(A) 28 (B) 21 (C) 40 (D) 30

Answer: C

b) A Sample Microeconomics Question

One of the reasons that the government discourages and regulates monopolies is that

- (A) producer surplus is lost and consumer surplus is gained. ✗
- (B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
- (C) monopoly firms do not engage in significant research and development. ✗
- (D) consumer surplus is lost with higher prices and lower levels of output. ✓

Microeconomics

Figure 2.4.11

Gopher, Chinchilla, and variants of PaLM have each posted state-of-the-art results on MMLU. The current top result on MMLU comes from Flan-PaLM, a Google model that reports an average score of 75.2% (Figure 2.4.12).

MMLU: Average Weighted Accuracy

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report

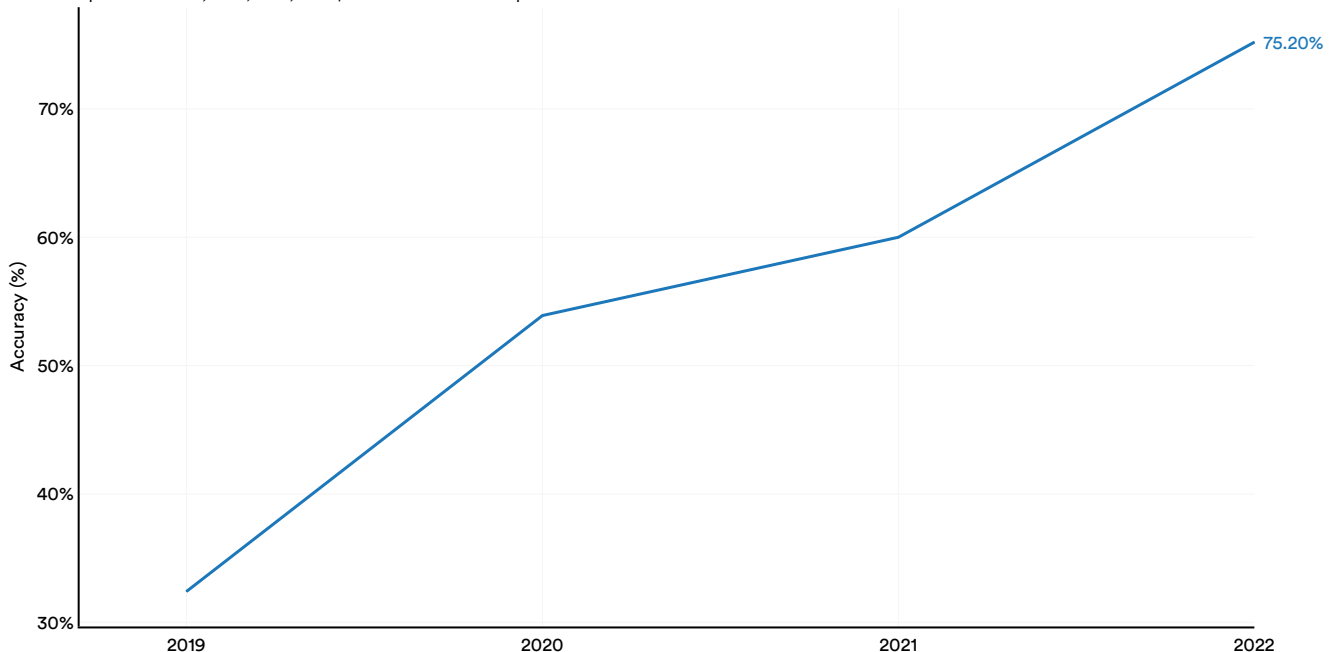


Figure 2.4.12

¹³ This criticism is more formally articulated in Hendrycks et al., 2021.



Machine Translation (MT)

Machine translation studies how well AI software can translate languages. In the last five years, machine translation has been dominated by neural networks which power current tools like DeepL and Google Translate.

Number of Commercially Available MT Systems

The popularity of AI-based machine translation is manifested in the number of commercial machine translation services on the market. Since 2017, the total number of independent machine translation services has increased six times (Figure 2.4.13).

Number of Independent Machine Translation Services

Source: Intento, 2022 | Chart: 2023 AI Index Report

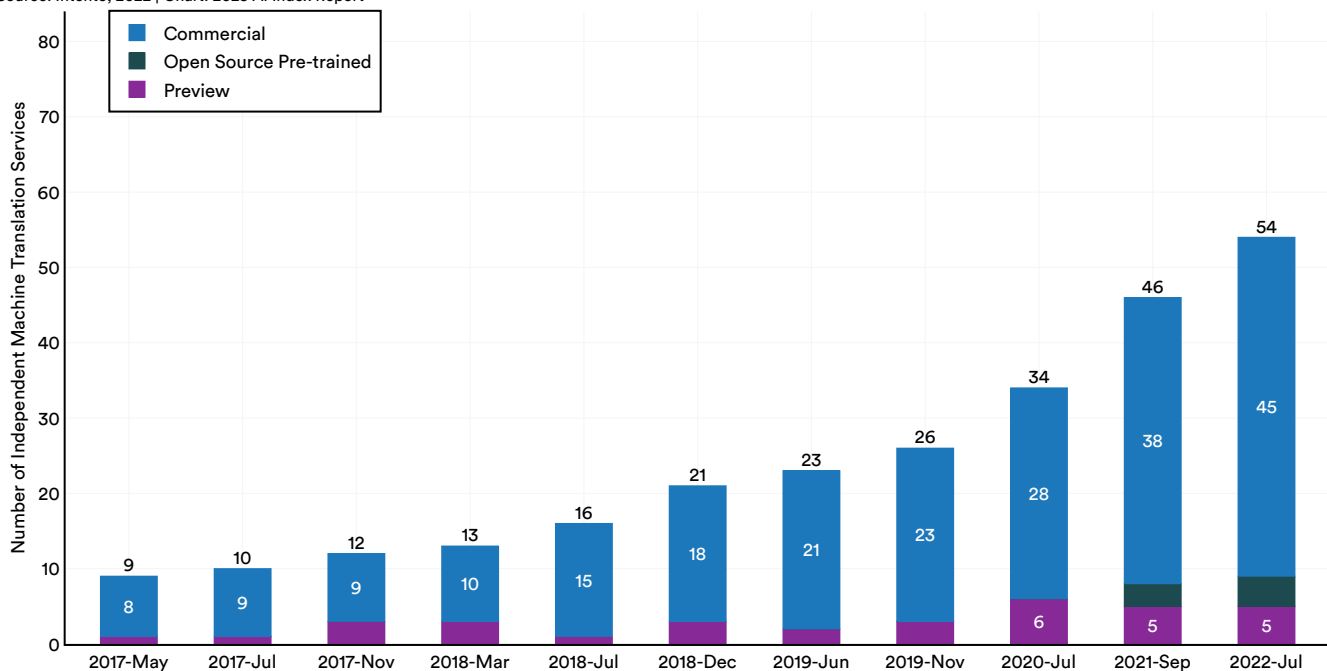


Figure 2.4.13



AI systems that work with human speech are usually tasked with converting spoken words into text and recognizing the individuals speaking.

2.5 Speech

Speech Recognition

Speech recognition is the ability of AI systems to identify spoken words and convert them into text. Speech recognition has progressed so much so that nowadays many computer programs or texting apps are equipped with dictation devices that can seamlessly transcribe speech into writing.

VoxCeleb

[VoxCeleb](#) is a large-scale audiovisual dataset of human speech for speaker recognition, which is the

task of matching certain speech with a particular individual. Over the years, the VoxCeleb dataset has been expanded; however, the data in this subsection tracks progress on the original dataset.

This year's top result on the original VoxCeleb dataset was posted by American researchers, whose model achieved an equal error rate of 0.1%, which represents a 0.28 percentage point decrease from the state-of-the-art result achieved by Chinese researchers in the previous year (Figure 2.5.1).

VoxCeleb: Equal Error Rate (EER)

Source: VoxCeleb, 2022 | Chart: 2023 AI Index Report

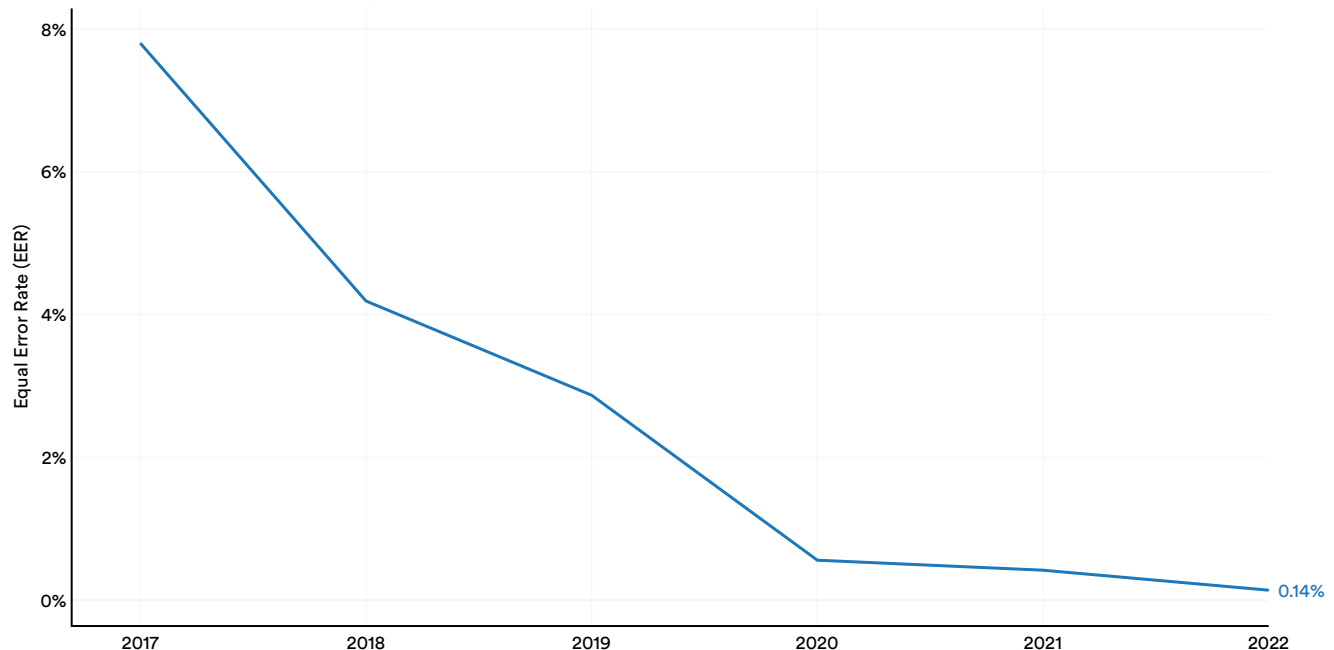


Figure 2.5.1

Narrative Highlight:

Whisper

One of the major themes in the last few years of AI progress has been the emergence of large language models that are trained on massive amounts of data and capable of executing a diverse range of tasks. In 2022, this idea of training on large data to achieve cross-domain performance arrived in the world of speech recognition with OpenAI's launch of [Whisper](#).

Whisper is a large-scale speech recognition model that was trained in a weakly supervised way on 700,000 hours of audio data. Whisper was capable of strong, although not state-of-the-art, performance on many speech recognition tasks in zero-shot settings.¹⁴ Whisper outperformed wav2vec 2.0 Large, another speech recognition model, across a wide range of popular English speech recognition benchmarks (Figure 2.5.2). Similarly, Whisper proved to be a better speech translator than many other leading AI translator models (Figure 2.5.3). Whisper also outperformed other commercial automated speech recognition systems and scored similarly to top human transcription services (Figure 2.5.4).¹⁵ Despite this impressive performance, there were still some speech tasks, like language identification, on which Whisper trailed state-of-the-art models (Figure 2.5.5).

wav2vec 2.0 Large (No LM) Vs. Whisper Large V2
Across Datasets

Source: Radford et al., 2022 | Chart: 2023 AI Index Report

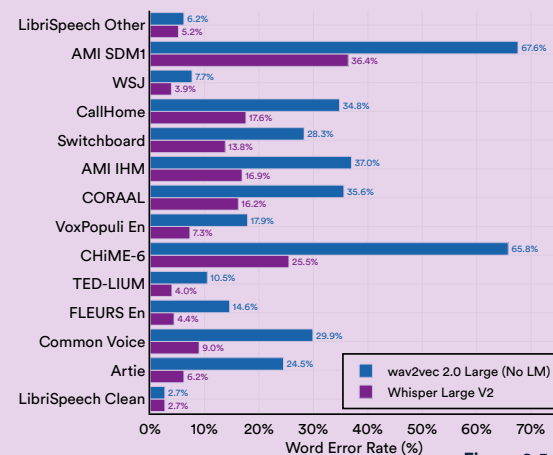


Figure 2.5.2

Notable Models on X→EN Subset of CoVoST 2

Source: Radford et al., 2022 | Chart: 2023 AI Index Report

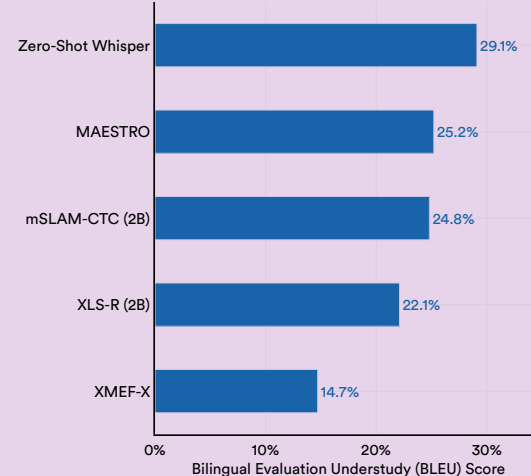


Figure 2.5.3

¹⁴ Zero-shot learning refers to the ability of an AI system to learn a particular task without being trained on that task.

¹⁵ Kincaid46 is a dataset of 46 audio files and transcripts that were published in the blog post, "Which automatic transcription service is the most accurate?—2018."

Narrative Highlight: Whisper (cont'd)

Notable Speech Transcription Services on Kincaid46

Source: Radford et al., 2022 | Chart: 2023 AI Index Report

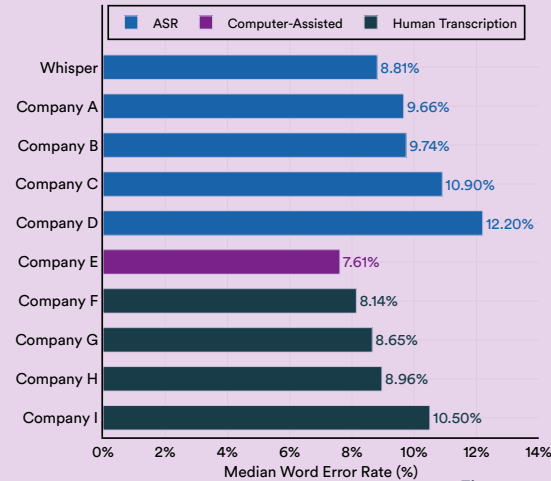


Figure 2.5.4

Notable Models on FLEURS: Language Identification Accuracy

Source: Radford et al., 2022 | Chart: 2023 AI Index Report

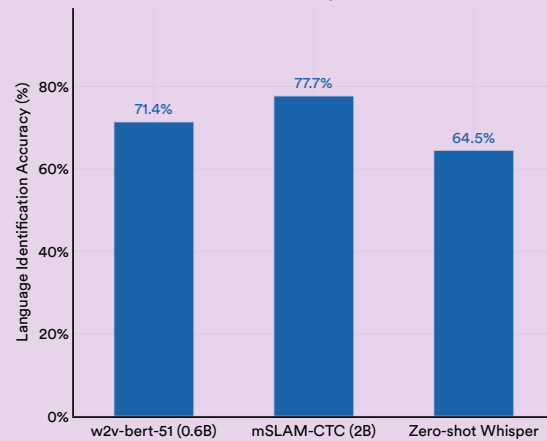


Figure 2.5.5

Whisper represents a breakthrough in state-of-the-art speech recognition systems. Traditionally, such systems were either pre-trained using supervised learning methods or pre-trained without supervision but required fine-tuning. Acquisition of data for supervised pre-training is time-consuming and costly. However, pre-training without supervision still requires further algorithmic specification to realize a desired objective like speech recognition. Algorithmic specification itself often requires a skilled practitioner. Whisper resolves these issues by demonstrating that a speech recognition system can perform well across a diverse range of tasks with massive amounts of unlabeled speech data.

In reinforcement learning, AI systems are trained to maximize performance on a given task by interactively learning from their prior actions. Systems are rewarded if they achieve a desired goal and punished if they fail.

2.6 Reinforcement Learning

Reinforcement Learning Environments

Reinforcement learning agents require environments, not datasets, to train: They must be trained in environments where they can experiment with various actions that will allow them to identify optimal game strategies.

Procgen

Procgen is a reinforcement learning environment introduced by OpenAI in 2019. It includes 16 procedurally generated video-game-like environments specifically designed to test the ability of reinforcement learning agents to learn generalizable skills (Figure 2.6.1). Performance on Procgen is measured in terms of mean-normalized score. Researchers typically train their systems on 200 million training runs and report an average score across the 16 Procgen games. The higher the system scores, the better the system.

The Different Environments in Procgen

Source: [OpenAI, 2019](#)

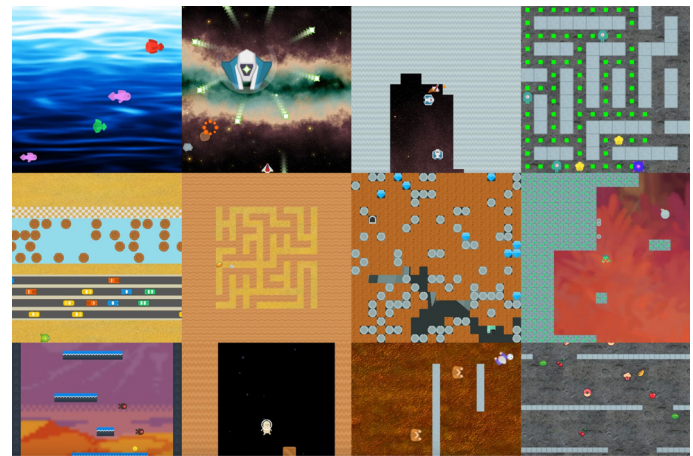


Figure 2.6.1



A team of industry and academic researchers from Korea posted the top score of 0.6 on Procgen in 2022 (Figure 2.6.2).

Procgen: Mean of Min-Max Normalized Score

Source: arXiv, 2022 | Chart: 2023 AI Index Report

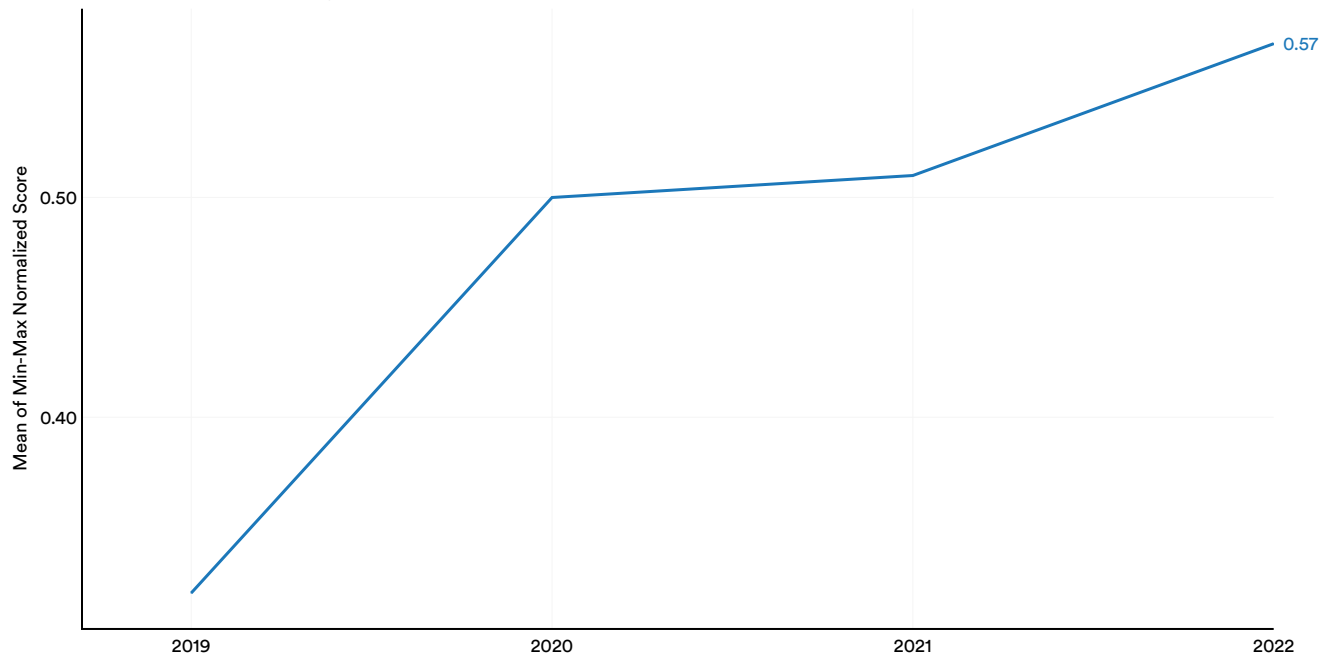


Figure 2.6.2

Narrative Highlight:

Benchmark Saturation

An emerging theme in this year's AI Index is the observed performance saturation across many popular technical performance benchmarks. Last year's AI Index Report observed a similar trend; however, benchmark saturation has been particularly pronounced this year. Figure 2.6.3 shows the relative improvement since the benchmark first launched (overall improvement) and relative improvement within the last year (YoY improvement) on AI technical benchmarks considered in this year's AI Index. The improvements are reported as percent changes.

For all but 7 of the benchmarks, the improvement registered is less than 5%. The median improvement within the last year is 4%, while the median improvement since launch is 42.4%.¹⁶ Moreover, this year the AI Index elected not to feature traditionally popular benchmarks like SQuAD1.1 and SQuAD2.0, as no new state-of-the-art results were posted. Moreover, the speed at which benchmark saturation is being reached is increasing. Researchers have responded to this increasing saturation by launching newer and more comprehensive benchmarking suites such as [BIG-bench](#) and [HELM](#).

Improvement Over Time on Select AI Index Technical Performance Benchmarks

Source: AI Index, 2022 | Chart: 2023 AI Index Report

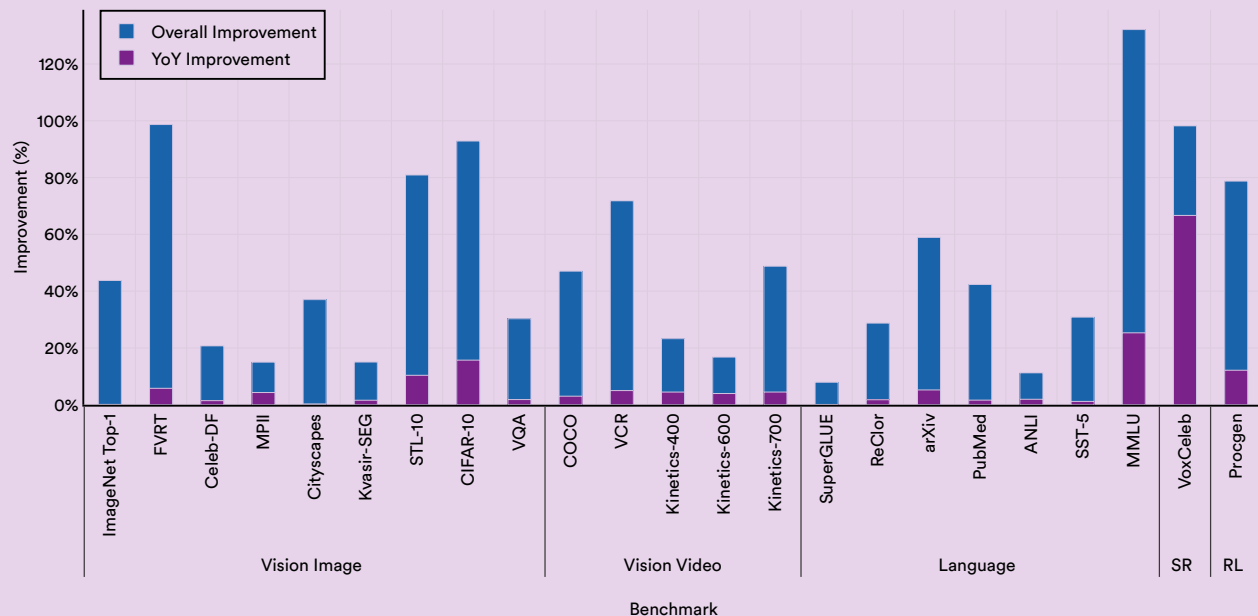


Figure 2.6.3

¹⁶ The improvements reviewed in this section are reported as relative change. Figure 2.6.3 should therefore not be used to conduct comparisons of improvements across benchmarks, as each benchmark has different parameters.



Deep learning AI algorithms are trained on GPUs or TPUs, which accelerate the training speed of AI systems. As AI systems process ever-larger datasets, it is crucial to monitor advancements in hardware capabilities.

2.7 Hardware

MLPerf Training

MLPerf is an AI training competition run by the ML Commons organization. In this challenge, participants train ML systems to execute various tasks using a common architecture. Entrants are then ranked on their absolute wall clock time, which is how long it takes for the system to train.

Last year, the AI Index observed that since the competition launched, training times for virtually

every AI skill category had significantly decreased. This year, this trend has continued, albeit at a slightly slower pace. Record-low training times were posted in the object detection, speech recognition, image segmentation, recommendation, image classification, and language processing categories (Figure 2.7.1). In categories like image classification and object detection, the top AI systems can now train roughly 32 times quicker than in 2018, when the competition first launched.

MLPerf Training Time of Top Systems by Task: Minutes

Source: MLPerf, 2022 | Chart: 2023 AI Index Report

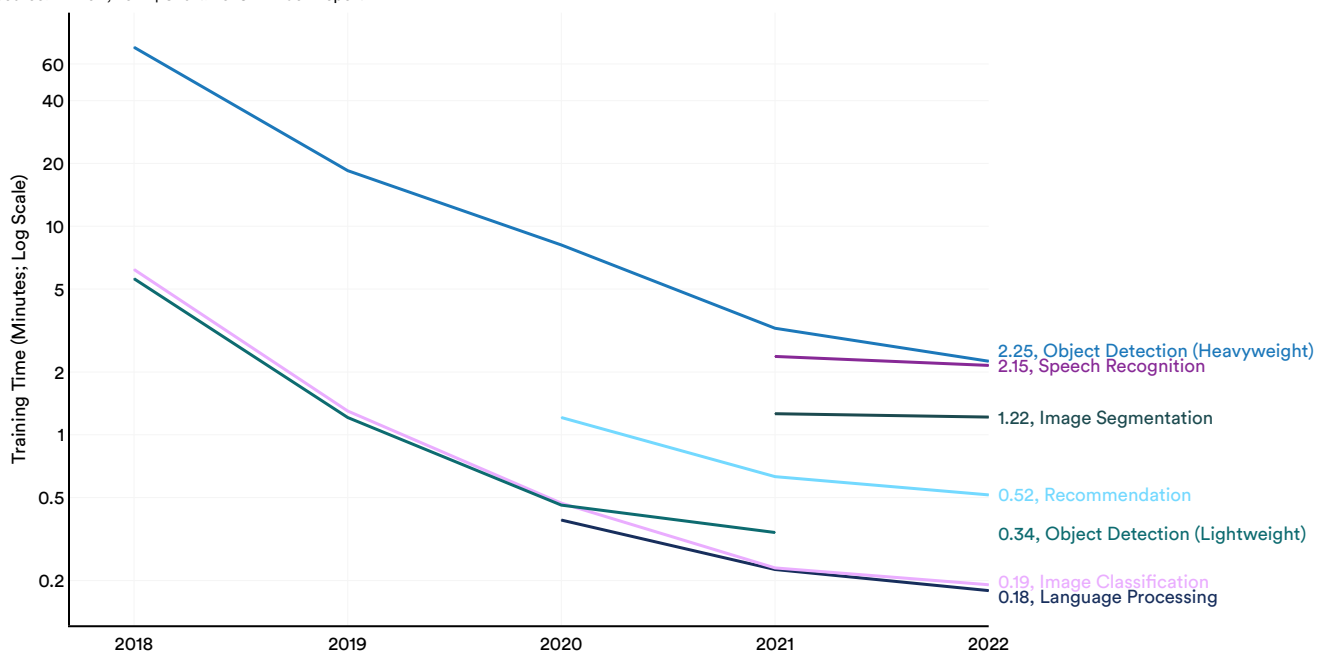


Figure 2.7.1



Data on the number of accelerators used by the hardware systems submitted to MLPerf also suggests that stronger hardware has been powering decreasing training times (Figure 2.7.2). Since the start of the MLPerf competition, the gap has grown

between the mean number of accelerators used by all entrants and the average accelerators used by the systems that post the top results.¹⁷ This gap suggests that having better hardware is essential to training the fastest systems.

MLPerf Hardware: Accelerators

Source: MLPerf, 2022 | Chart: 2023 AI Index Report

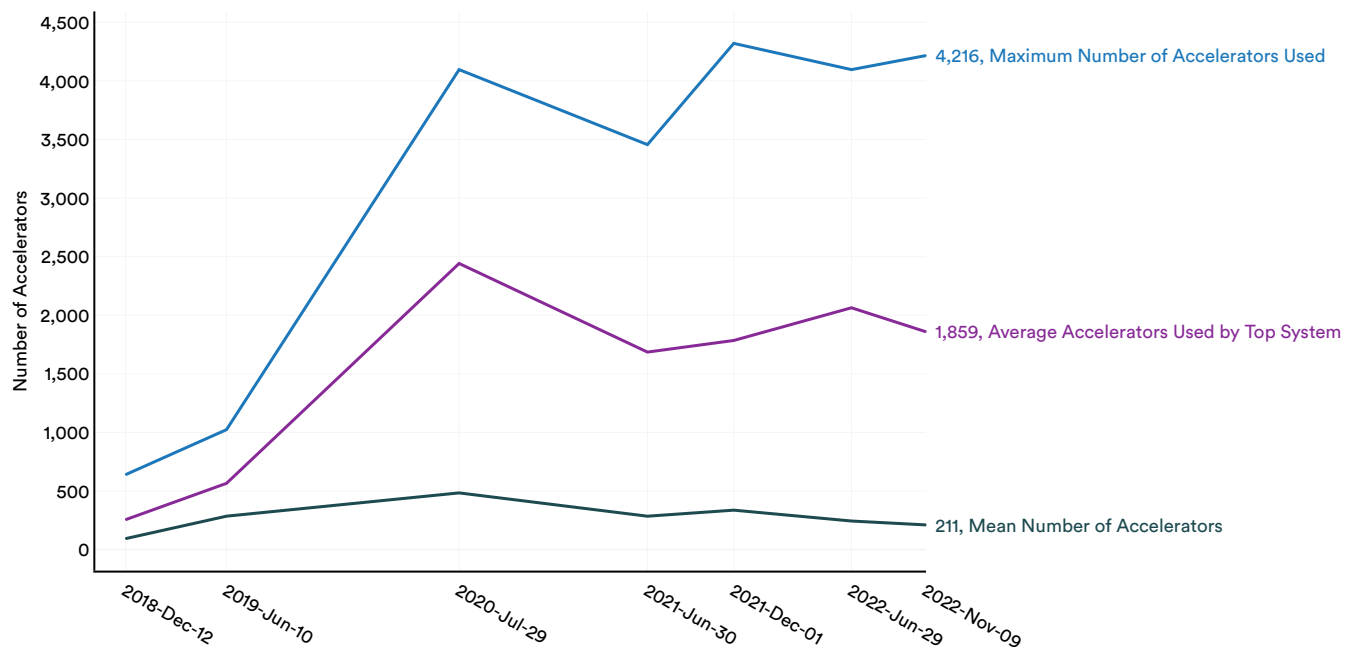


Figure 2.7.2

¹⁷ An accelerator, like a GPU or TPU, is a chip that is chiefly used for the machine learning component of a training run.



MLPerf Inference

In deploying AI, inference is the step where trained AI systems generate predictions, e.g. classifying objects.

In 2020, ML Commons introduced MLPerf Inference, a performance benchmarking suite that measures how fast a trained AI system can process inputs and produce inferences. The MLPerf Inference suite tracks the throughput of AI systems, measured in samples per second or queries per second.¹⁸

Figures 2.7.3 to 2.7.6 plot the throughput of the state-of-the-art submissions on MLPerf Inference across four skill categories: image classification, language processing, recommendation, and speech recognition. The number of inferences generated by the top-performing AI systems has significantly increased since the first iteration of the competition in 2020. For example, the number of offline samples generated by the top image classifiers and language processors have more than doubled since 2020, while those for recommendation systems have increased by roughly 23%.

MLPerf Best-Performing Hardware for Image Classification: Offline and Server Scenario

Source: MLPerf, 2022 | Chart: 2023 AI Index Report

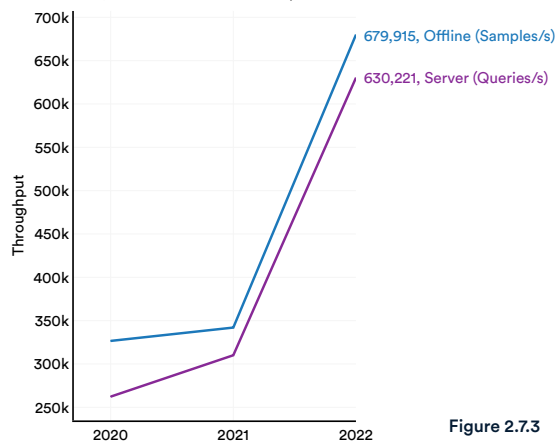


Figure 2.7.3

MLPerf Best-Performing Hardware for Language Processing: Offline and Server Scenario

Source: MLPerf, 2022 | Chart: 2023 AI Index Report

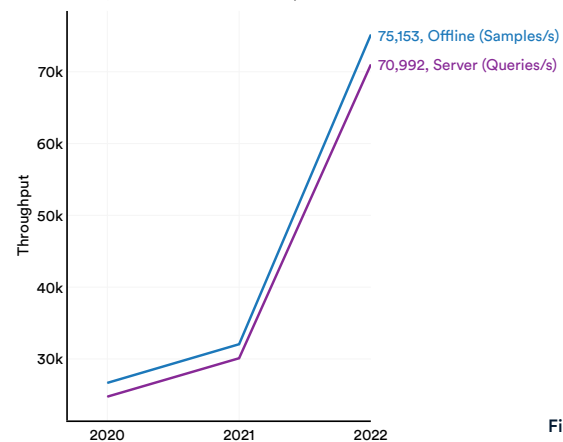


Figure 2.7.4

MLPerf Best-Performing Hardware for Recommendation: Offline and Server Scenario

Source: MLPerf, 2022 | Chart: 2023 AI Index Report

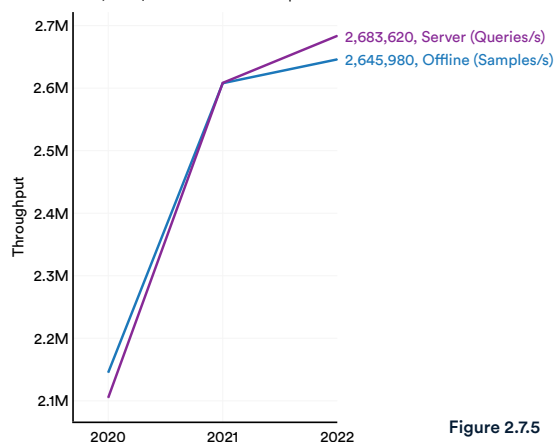


Figure 2.7.5

MLPerf Best-Performing Hardware for Speech Recognition: Offline and Server Scenario

Source: MLPerf, 2022 | Chart: 2023 AI Index Report

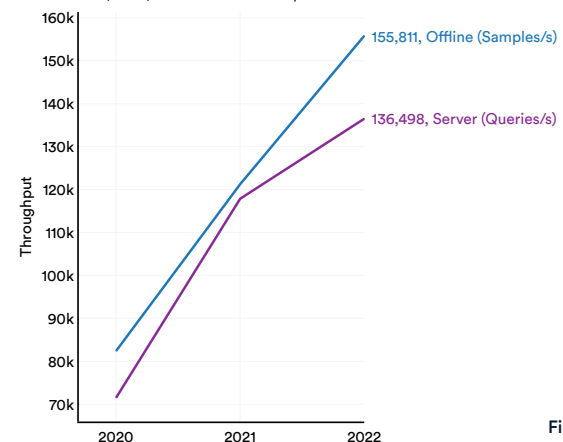


Figure 2.7.6

¹⁸ The following [blog post](#) from Dell Technologies offers a good distinction between offline and server samples: "Offline—one query with all samples is sent to the system under test (SUT). The SUT can send the results back once or multiple times in any order. The performance metric is samples per second. Server—the queries are sent to the SUT following a Poisson distribution (to model real-world random events). One query has one sample. The performance metric is queries per second (QPS) within the latency bound."

Trends in GPUs: Performance and Price

This year, the AI Index built on [work](#) previously done by the research collective Epoch and analyzed trends over time in GPU performance and price.¹⁹

Figure 2.7.7 showcases the FP32 (single precision) performance FLOP/s of different GPUs released from 2003 to 2022. FLOP/s stands for “Floating Point Operations per second” and is a measure of

the performance of a computational device. The higher the FLOP/s, the better the hardware.

Figure 2.7.8 showcases the median single performance of new GPUs by release date, which continues to rise year over year. Since 2021, the median FLOP/s speed has nearly tripled, and since 2003 it has increased roughly 7,000 times.

FP32 (Single Precision) Performance (FLOP/s) by Hardware Release Date, 2003–22

Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

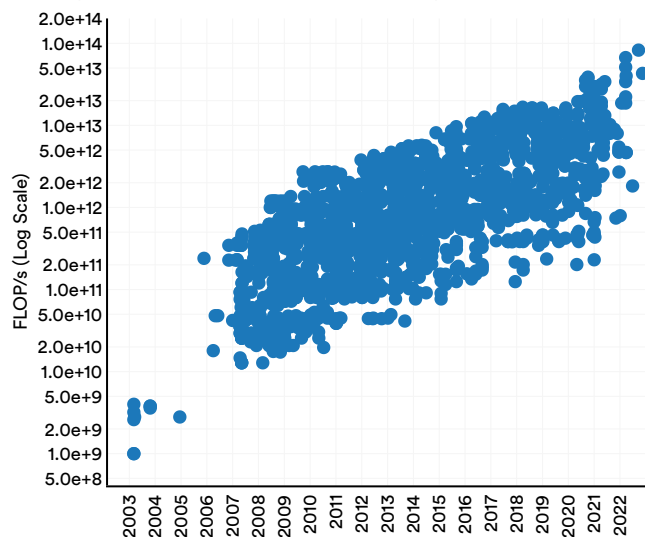


Figure 2.7.7

Median FP32 (Single Precision) Performance (FLOP/s), 2003–22

Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

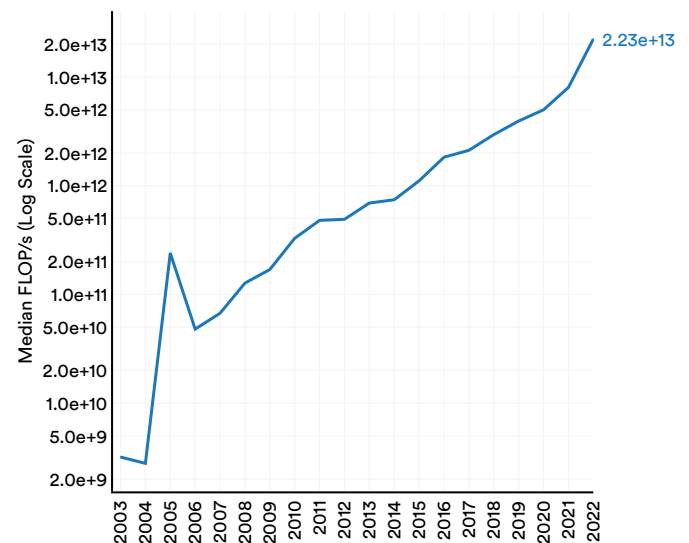


Figure 2.7.8

¹⁹ The Appendix fully delineates both the methodology of this approach and the unique ways in which AI Index research built upon the existing Epoch research.

Finally, figures 2.7.9 and 2.7.10 consider GPU trends in terms of FLOP/s per U.S. Dollar.²⁰ This statistic considers whether the underlying performance of GPUs is increasing relative to their changing costs. As evidenced most clearly in Figure 2.7.10, the price–performance of GPUs is rapidly increasing. The median FLOP/s per U.S. Dollar of GPUs in

2022 is 1.4 times greater than it was in 2021 and 5600 times greater than in 2003, showing a doubling in performance every 1.5 years. As noted in [similar analyses](#), improvements in the price–performance of AI hardware has facilitated increasingly larger training runs and encouraged the scaling of large AI models.

FP32 (Single Precision) Performance (FLOP/s) per U.S. Dollar by Hardware Release Date, 2003–22

Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

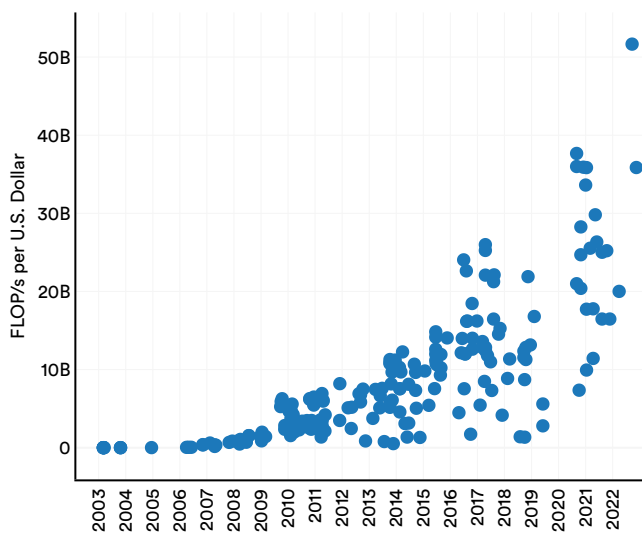


Figure 2.7.9

Median FP32 (Single Precision) Performance (FLOP/s) per U.S. Dollar, 2003–22

Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

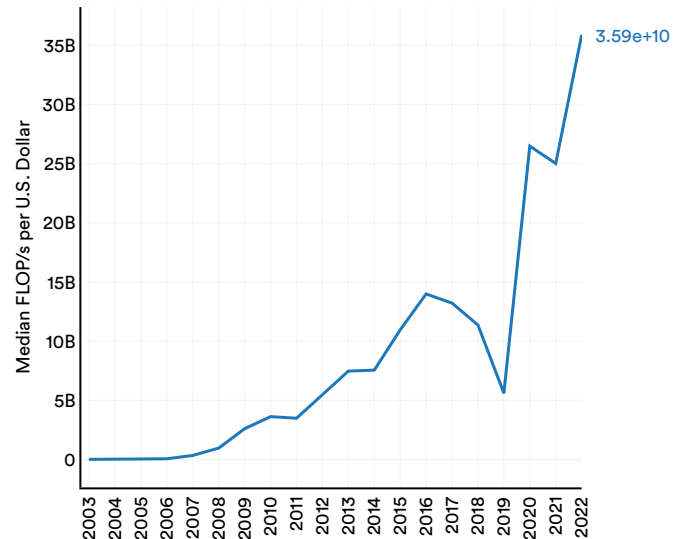


Figure 2.7.10

²⁰ The data in figures 2.7.9 and 2.7.10 has been adjusted for inflation. The exact details of the adjustment are outlined in greater detail in the Appendix.



There have been mounting concerns about the environmental impact of computational resources and the energy required for AI training and inference. Although there is no standard benchmark for tracking the carbon intensity of AI systems, this subsection synthesizes the findings of different researchers who are exploring the link between AI and the environment. Conducting research on the environmental effects of AI was challenging as there are wildly varying estimates, the validity of which have not yet been definitively established. To that end, the AI Index focuses on research from a recent paper by [Luccioni et al., 2022](#). As AI models continue growing in size and become more universally deployed, it will be increasingly important for the AI research community to consciously monitor the effect AI systems have on the environment.

2.8 Environment

Environmental Impact of Select Large Language Models

Many factors determine the amount of carbon emissions emitted by AI systems, including the number of parameters in a model, the power usage effectiveness of a data center, and the grid carbon intensity. Power Usage Effectiveness (PUE) is a metric used to evaluate the energy efficiency of data centers. It is the ratio of the total amount of energy used by a computer data center facility, including air conditioning, to the energy delivered to computing equipment. The higher the PUE, the less efficient the data center. Figure 2.8.1 shows how these factors compare across four large language models: GPT-3, Gopher, OPT, and BLOOM. It is

challenging to directly compare the carbon footprint of these models, as the accounting methodologies for reporting carbon emissions are not standardized.

Of the four language models being compared, GPT-3 released the most carbon, 1.4 times more than Gopher, 7.2 times more than OPT, and 20.1 times more than BLOOM.

Figure 2.8.2 relativizes the carbon-emission estimates to real-life examples. For instance, BLOOM's training run emitted 1.4 times more carbon than the average American uses in one year and 25 times that of flying one passenger round trip from New York to San Francisco. BLOOM's training consumed enough energy to power the average American home for 41 years.²¹

Environmental Impact of Select Machine Learning Models, 2022

Source: Luccioni et al., 2022 | Table: 2023 AI Index Report

Model	Number of Parameters	Datacenter PUE	Grid Carbon Intensity	Power Consumption	CO2 Equivalent Emissions	CO2 Equivalent Emissions x PUE
Gopher	280B	1.08	330 gCO2eq/kWh	1,066 MWh	352 tonnes	380 tonnes
BLOOM	176B	1.20	57 gCO2eq/kWh	433 MWh	25 tonnes	30 tonnes
GPT-3	175B	1.10	429 gCO2eq/kWh	1,287 MWh	502 tonnes	552 tonnes
OPT	175B	1.09	231 gCO2eq/kWh	324 MWh	70 tonnes	76.3 tonnes

Figure 2.8.1

²¹ The U.S. Energy Information Administration estimates that in 2021, the average annual electricity consumption of a U.S. residential utility customer was 10,632 kilowatt hours (kWh).



CO2 Equivalent Emissions (Tonnes) by Selected Machine Learning Models and Real Life Examples, 2022

Source: Luccioni et al., 2022; Strubell et al., 2019 | Chart: 2023 AI Index Report

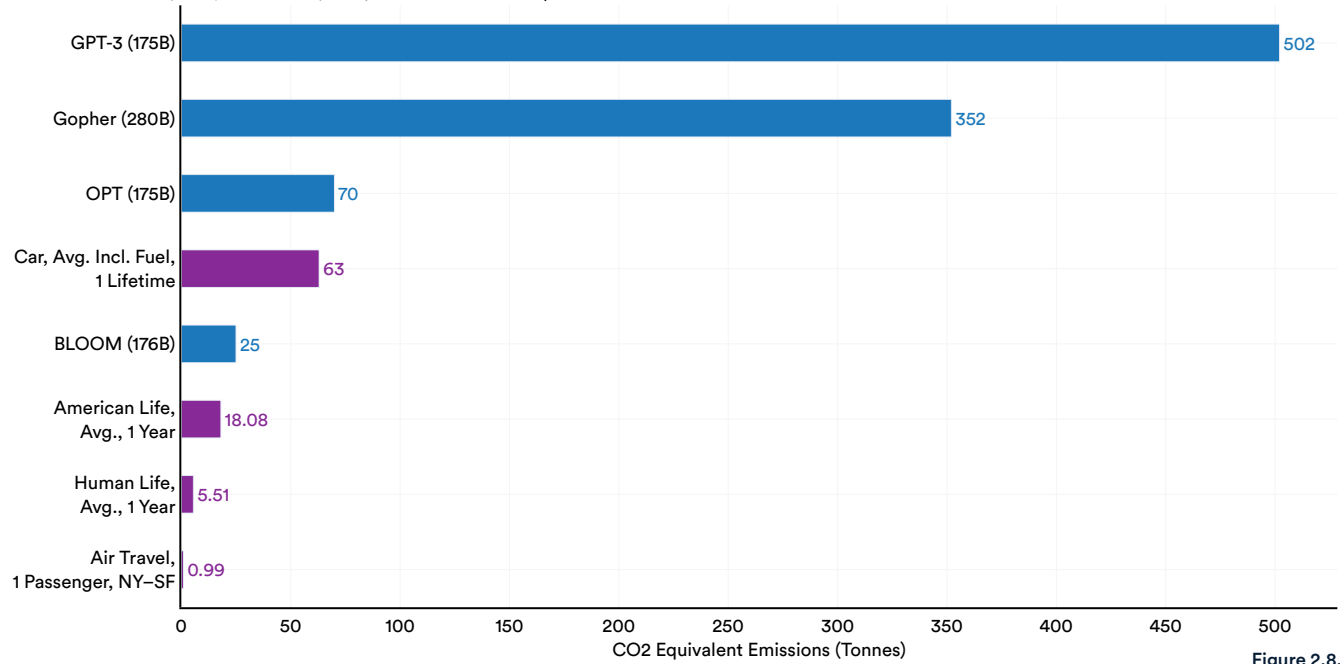


Figure 2.8.2

Narrative Highlight:

Using AI to Optimize Energy Usage

Training AI systems can be incredibly energy intensive. At the same time, recent research suggests that AI systems can be used to optimize energy consumption. In 2022, [DeepMind](#) released the results of a 2021 experiment in which it trained a reinforcement learning agent called BCOOLER (BVE-based COntstrained Optimization Learner with Ensemble Regularization) to optimize cooling procedures for Google's data centers.

Figure 2.8.3 presents the energy-saving results from one particular BCOOLER experiment. At the end of the three-month experiment, BCOOLER achieved roughly 12.7% energy savings. BCOOLER was able to achieve these savings while maintaining the cooling comfort levels that the building managers preferred.

Energy Savings Results Over Time for Select BCOOLER Experiment

Source: Luo et al., 2022 | Chart: 2023 AI Index Report

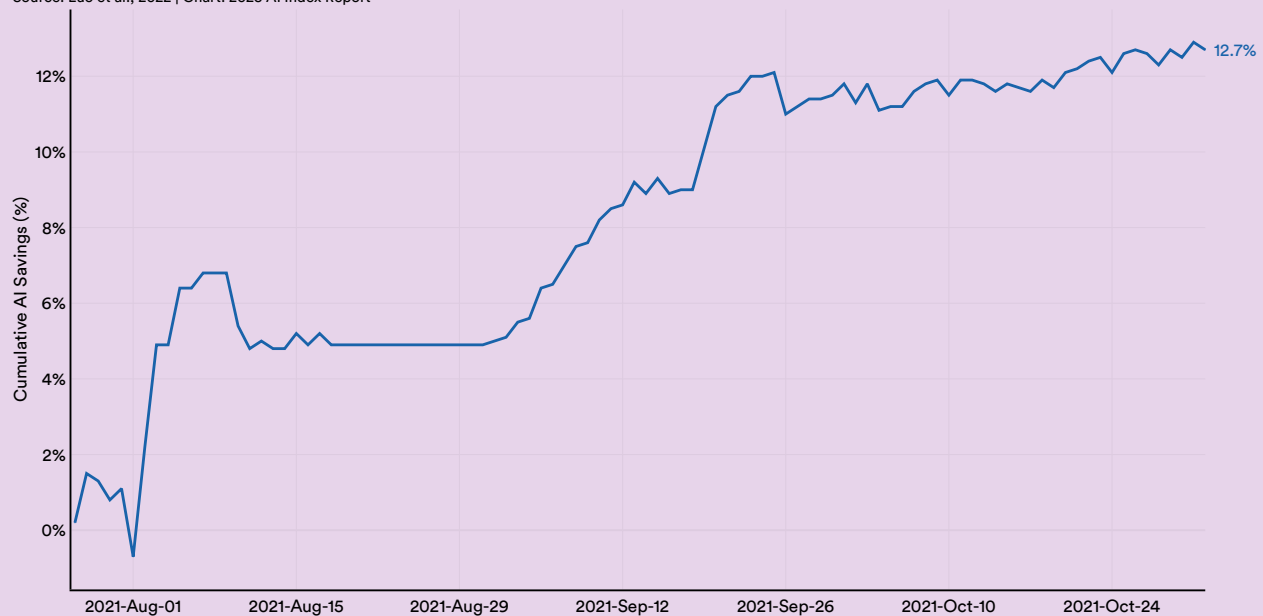


Figure 2.8.3

2022 was a groundbreaking year for AI in science. This subsection looks at some meaningful ways in which AI has recently been used to accelerate scientific discovery.

2.9 AI for Science

Accelerating Fusion Science Through Learned Plasma Control

Nuclear fusion could generate clean energy by fusing hydrogen. A common approach to achieving nuclear fusion is using a tokamak, a machine which controls and contains the heated hydrogen plasma (Figure 2.9.1). However, the plasmas produced in these machines are unstable and necessitate constant monitoring. In 2022, researchers at DeepMind developed a reinforcement learning algorithm to discover optimal tokamak management procedures.

Photos of the Variable Configuration Tokamak (TCV) at EPFL

Source: [DeepMind, 2022](#)

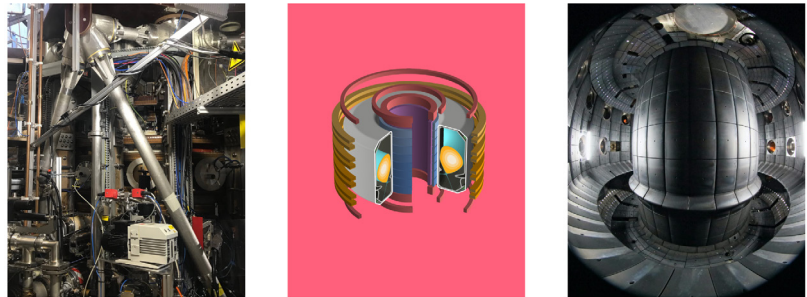


Figure 2.9.1

Discovering Novel Algorithms for Matrix Manipulation With AlphaTensor

Matrix multiplication is a simple algebraic operation that is essential to many computations, including neural networks and scientific computing (Figure 2.9.2). The classic algorithm to multiply two 2×2 matrices takes $2^3 = 8$ multiplications. Strassen discovered 50 years ago how to reduce this to 7, and generally how to multiply two $n \times n$ matrices in $O(n^{\log(7)})$ operations. DeepMind's AlphaTensor uses Reinforcement Learning to improve on state-of-the-art algorithms for many matrix sizes,

A Demonstration of AlphaTensor's Matrix Manipulation Process

Source: [Fawzi et al., 2022](#)

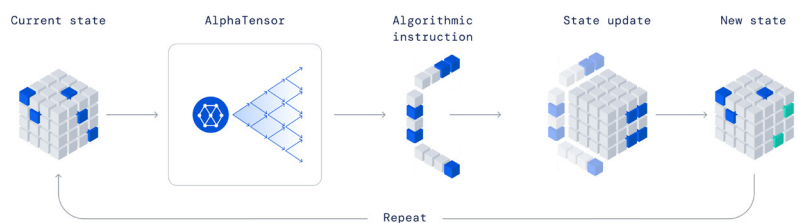


Figure 2.9.2

including 4×4 matrices over the integers $[0,1]$. It also matches state-of-the-art performance on several other matrix sizes, including 4×4 over the integers. It does this by searching through large numbers of possible algorithms, and evaluating them over real computer architectures.

Designing Arithmetic Circuits With Deep Reinforcement Learning

This year, a team at Nvidia discovered a novel approach to improving the chips that power AI systems: Use AI systems to design better chips. They were able to train a reinforcement learning agent to design chip circuits that are smaller, faster, and more efficient than the circuits designed by electronic design automation tools (EDAs). One of Nvidia's latest categories of chips, the Hopper GPU architecture, has over 13,000 instances of AI-designed circuits. Figure 2.9.3 shows a 64-bit adder circuit designed by Nvidia's PrefixRL AI agent (on the left) which is 25% smaller while being just as fast and functional as those designed by the state-of-the-art EDA tools.

A Juxtaposition of Nvidia Circuits Designed by PrefixRL Vs. EDA Tools

Source: Roy et al., 2022

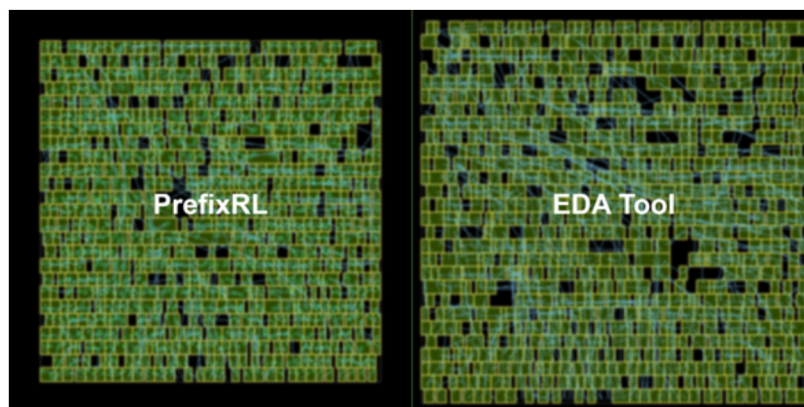


Figure 2.9.3

Unlocking de Novo Antibody Design With Generative AI

Antibody discovery, which is referred to as de novo antibody discovery, typically requires immense amounts of time and resources. Traditional methods for de novo discovery offer little control over the outputs, so that proposed antibodies are often suboptimal. To that end, a team of researchers turned to generative AI models to create antibodies in a zero-shot fashion, where antibodies are created with one round of model generation without further optimizations (Figure 2.9.4). These AI-generated antibodies are also robust. The fact that generative AI can create new antibodies has the potential to accelerate drug discovery.

Zero-Shot Generative AI for de Novo Antibody Design

Source: Shانهsazzadeh et al., 2023

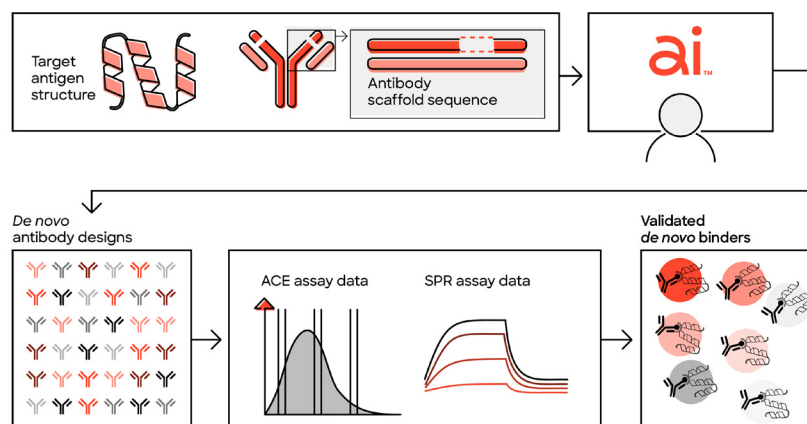


Figure 2.9.4



Appendix

ImageNet

Data on ImageNet accuracy was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (top-1 accuracy) correspond to the result reported in the most recent version of each paper. Learn more about the LSVRC ImageNet competition and the [ImageNet dataset](#).

To highlight progress on top-1 accuracy without the use of extra training data, scores were taken from the following papers:

[Aggregated Residual Transformations for Deep Neural Networks](#)
[Exploring the Limits of Weakly Supervised Pretraining](#)
[Fixing the Train-Test Resolution Discrepancy: FixEfficientNet](#)
[ImageNet Classification With Deep Convolutional Neural Networks](#)
[PeCo: Perceptual Codebook for BERT](#)
[Pre-training of Vision Transformers](#)
[Progressive Neural Architecture Search](#)
[Rethinking the Inception Architecture for Computer Vision](#)
[Self-Training With Noisy Student Improves ImageNet Classification](#)
[Some Improvements on Deep Convolutional Neural Network Based Image Classification](#)

[Very Deep Convolutional Networks for Large-Scale Image Recognition](#)

[ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond](#)

To highlight progress on top-1 accuracy with the use of extra training data, scores were taken from the following papers:

[Big Transfer \(BiT\): General Visual Representation Learning](#)

[CoAtNet: Marrying Convolution and Attention for All Data Sizes](#)

[CoCa: Contrastive Captioners Are Image-Text Foundation Models](#)

[Meta Pseudo Labels](#)

National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT)

Data on NIST FRVT 1:1 verification accuracy by dataset was obtained from the [FRVT 1:1 verification leaderboard](#).

Celeb-DF

Data on Celeb-DF AUC was retrieved through a detailed arXiv literature review. The reported dates correspond to the year in which a paper was first published to arXiv or a method was introduced. With Celeb-DF, recent researchers have tested previously existing deepfake detection methodologies. The year in which a method was introduced, even if it was subsequently tested, is the year in which it is included in the report. The reported results (AUC) correspond to the result reported in the most recent version of each paper. Details on the Celeb-DF benchmark can be found in the [Celeb-DF paper](#).

To highlight progress on Celeb-DF, scores were taken from the following papers:

[Deepfake Detection via Joint Unsupervised Reconstruction and Supervised Classification](#)

[Exposing Deepfake Videos by Detecting Face Warping Artifacts](#)

[Face X-Ray for More General Face Forgery Detection](#)
[FaceForensics++: Learning to Detect Manipulated Facial Images](#)

[Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain](#)

MPII

Data on MPII percentage of correct keypoints (PCK) was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (PCK) correspond to the result reported in the most recent

version of each paper. Details on the MPII benchmark can be found in the [MPII paper](#) and [MPII dataset](#).

To highlight progress on percentage of correct keypoints without the use of extra training data, scores were taken from the following papers:

[Bottom-Up and Top-Down Reasoning With Hierarchical Rectified Gaussians](#)

[Cascade Feature Aggregation for Human Pose Estimation](#)

[Deeply Learned Compositional Models for Human Pose Estimation](#)

[Efficient Object Localization Using Convolutional Networks](#)

[Learning Feature Pyramids for Human Pose Estimation](#)

[Stacked Hourglass Networks for Human Pose Estimation](#)

[Toward Fast and Accurate Human Pose Estimation via Soft-Gated Skip Connections](#)

[ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation](#)

Cityscapes Challenge, Pixel-Level Semantic Labeling Task

Data on the Cityscapes challenge, pixel-level semantic labeling task mean intersection-over-union (mIoU) was taken from the Cityscapes dataset, specifically their [pixel-level semantic labeling leaderboard](#). More details about the Cityscapes dataset and other corresponding semantic segmentation challenges can be accessed at the [Cityscapes dataset webpage](#).

Kvasir-SEG

Data on Kvasir-SEG mean dice was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (mean dice) correspond to the result reported in the most recent version of each paper. Details on the Kvasir-SEG benchmark can be found in the [Kvasir-SEG paper](#).

To highlight progress on Kvasir-SEG, scores were taken from the following papers:

[GMSRF-Net: An Improved Generalizability With Global Multi-Scale Residual Fusion Network for Polyp Segmentation](#)

[PraNet: Parallel Reverse Attention Network for Polyp Segmentation](#)

[ResUNet++: An Advanced Architecture for Medical Image Segmentation](#)

[Spatially Exclusive Pasting: A General Data Augmentation for the Polyp Segmentation](#)

Common Object in Context (COCO)

Data on COCO mean average precision (mAP50) was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (mAP50) correspond to the result reported in the most recent version of each paper. Details on the COCO benchmark can be found in the [COCO paper](#).

To highlight progress on COCO, scores were taken from the following papers:

[An Analysis of Scale Invariance in Object Detection-SNIP](#)

[CBNet: A Novel Composite Backbone Network Architecture for Object Detection](#)

[Deformable ConvNets v2: More Deformable, Better Results](#)

[DetectoRS: Detecting Objects With Recursive Feature Pyramid and Switchable Atrous Convolution](#)

[EVA: Exploring the Limits of Masked Visual Representation Learning at Scale](#)

[Grounded Language-Image Pre-training](#)

[Inside-Outside Net: Detecting Objects in Context With Skip Pooling and Recurrent Neural Networks](#)

CIFAR-10

Data on CIFAR-10 FID scores was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (FID score) correspond to the result reported in the most recent version of each paper. Details on the CIFAR-10 benchmark can be found in the [CIFAR-10 paper](#).

To highlight progress on CIFAR-10, scores were taken from the following papers:

[GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium](#)

[Large Scale GAN Training for High Fidelity Natural Image Synthesis](#)

[Refining Generative Process With Discriminator Guidance in Score-Based Diffusion Models](#)

[Score-Based Generative Modeling in Latent Space](#)

[Score-Based Generative Modeling Through Stochastic Differential Equations](#)

[Self-Supervised GAN: Analysis and Improvement With Multi-Class Minimax Game](#)

STL-10

Data on STL-10 FID scores was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (FID score) correspond to the result reported in the most recent version of each paper. Details on the STL-10 benchmark can be found in the [STL-10 paper](#).

To highlight progress on STL-10, scores were taken from the following papers:

[DEGAS: Differentiable Efficient Generator Search](#)

[Diffusion-GAN: Training GANs With Diffusion](#)

[Discriminator Contrastive Divergence: Semi-Amortized Generative Modeling by Exploring Energy of the Discriminator](#)

[Dist-GAN: An Improved GAN Using Distance Constraints](#)

[Soft Truncation: A Universal Training Technique of Score-Based Diffusion Model for High Precision Score Estimation](#)

Text-to-Image Models on MS-COCO 256 × 256 FID-30K

Data on MS-COCO 256 x 256 FID 30K for Text-to-Image Models was retrieved from the paper [Saharia et al., 2022](#).

Visual Question Answering (VQA)

Data on VQA accuracy was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#).

The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (accuracy) correspond to the result reported in the most recent version of each paper. Human-level performance is taken from the 2021 [VQA challenge](#).

To highlight progress on VQA accuracy without the use of extra training data, scores were taken from the following papers:

[Bilinear Attention Networks](#)

[Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding](#)

[Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks](#)

[PaLI: A Jointly-Scaled Multilingual Language-Image Model](#)

[Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge](#)

[UNITER: UNiversal Image-TExt Representation Learning](#)

[VLMo: Unified Vision-Language Pre-training With Mixture-of-Modality-Experts](#)

BEiT-3 Vs. Previous SOTA

Data on BEiT-3 and Previous SOTA was retrieved from the paper [Wang et al., 2022](#).

Visual Commonsense Reasoning (VCR)

Data on VCR Q->AR score was taken from [VCR leaderboard](#); the VCR leaderboard webpage further delineates the methodology behind the VCR challenge. Human performance on VCR is taken from [Zellers et al., 2018](#). Details on the VCR benchmark can be found in the [VCR paper](#).

Kinetics-400, Kinetics-600, and Kinetics-700

Data on Kinetics-400, Kinetics-600, and Kinetics-700 accuracy was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on Papers With Code ([Kinetics-400](#), [Kinetics-600](#), and [Kinetics-700](#)). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (top-1 accuracy) correspond to the result reported in the most recent version of each paper. Details on the Kinetics-400 benchmark can be found in the [Kinetics-400 paper](#). Details on the Kinetics-600 benchmark can be found in the [Kinetics-600 paper](#). Details on the Kinetics-700 benchmark can be found in the [Kinetics-700 paper](#).

To highlight progress on Kinetics-400, scores were taken from the following papers:

[Co-training Transformer With Videos and Images Improves Action Recognition](#)

[InternVideo: General Video Foundation Models via Generative and Discriminative Learning](#)

[Large-Scale Weakly-Supervised Pre-training for Video Action Recognition](#)

[Non-Local Neural Networks](#)

[Omni-Sourced Webly-Supervised Learning for Video Recognition](#)

[SlowFast Networks for Video Recognition](#)

[Temporal Segment Networks: Towards Good Practices for Deep Action Recognition](#)

To highlight progress on Kinetics-600, scores were taken from the following papers:

[Learning Spatio-Temporal Representation With Local and Global Diffusion](#)

[Masked Feature Prediction for Self-Supervised Visual Pre-training](#)

[PERF-Net: Pose Empowered RGB-Flow Net](#)

[Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-Offs in Video Classification](#)

[Rethinking Video ViTs: Sparse Video Tubes for Joint Image and Video Learning](#)

[SlowFast Networks for Video Recognition](#)

To highlight progress on Kinetics-700, scores were taken from the following papers:

[InternVideo: General Video Foundation Models via Generative and Discriminative Learning](#)

[Learn to Cycle: Time-Consistent Feature Discovery for Action Recognition](#)

[Masked Feature Prediction for Self-Supervised Visual Pre-training](#)

Text-to-Video Models on UCF-101

Data on UCF-101 Inception Score (IS) for Text-to-Video Models was retrieved from the paper [Hong et al., 2022](#), and [Singer et al., 2022](#).

SuperGLUE

The SuperGLUE benchmark data was pulled from the [SuperGLUE leaderboard](#). Details about the SuperGLUE benchmark are in the [SuperGLUE](#) paper and [SuperGLUE software toolkit](#). The tasks and evaluation metrics for SuperGLUE are:

NAME	IDENTIFIER	METRIC
Broadcoverage Diagnostics	AX-b	Matthew's Carr
CommitmentBank	CB	Avg. F1/Accuracy
Choice of Plausible Alternatives	COPA	Accuracy
Multi-Sentence Reading Comprehension	MultiRC	F1a/EM
Recognizing Textual Entailment	RTE	Accuracy
Words in Context	WiC	Accuracy
The Winograd Schema Challenge	WSC	Accuracy
BoolQ	BoolQ	Accuracy
Reading Comprehension With Commonsense Reasoning	ReCoRD	F1/Accuracy
Winogender Schema Diagnostic	AX-g	Gender Parity/Accuracy

Reading Comprehension Dataset Requiring Logical Reasoning (ReClor)

Data on ReClor accuracy was retrieved through the [ReClor leaderboard](#) cross-referenced by technical progress reported on [Papers With Code](#). Details about the ReClor benchmark can be found in the [ReClor paper](#).



Narrative Highlight: Just How Much Better Have Language Models Become?

Complete answer outputted by GPT-2 to the AI Index prompt request.

Model prompt >>> **Explain to me the major accomplishments of Theodore Roosevelt's presidency.**

(The most notable is the establishment of a national telephone and mail exchange company.)

But the only things that will change with a president who knows how to operate a business is his understanding of the American system of government. The public knows all about our problems, yet at least half of the people who need it won't know anything about ours. We can no longer have this. Our system of government is the only

one in which you really need a public relations firm to work hard.

The system of governmental control, while working, is not always effective in ensuring that the very things that our economy needs, do not always be at the top of everybody's list. So I wanted to know how successful the system you employ for managing the public's needs is in helping the American economy grow, how efficient the service our economy provides is being made, and how it can better serve the needs of a growing number of Americans.

I wanted to know.

This was before I ever hired my own lawyer to work on the case. This was before I ever hired my own law firm to handle government investigations. For my first major job, at least, I had to convince two top aides to get

paid \$90,000 for the role, and before I even started on the case, I had to convince one top staffer to get paid even \$100,000 to have an interview with my lead prosecutor.

I wanted to know in advance: What percentage of his work was actually done as a private citizen, and how efficiently should he handle it for his clients? And what advice would he give me should I end up in the public eye?

At no time, my job as the lead prosecutor on this case required me to put forth any of those two or five requests. Most people on the case, though, did. One person asked me what the case's real value was. That didn't stop me from telling it to one of my top managers. And my top manager never even mentioned it.

In the end, it was a two-week, \$40 million round-trip tour of the government's business, with the best lawyer possible. In exchange, we got a chance to hear his last public comment in this case, his last comment to me.

There was one more point — the end point of the case.

When it came up for consideration in October of 1993, I had the choice to give it three years, or six. And I chose six.

Large Language Models on the Blocksworld Domain

Data on Blocksworld domain for large language models was retrieved from the paper [Valmeekam et al., 2022](#).

arXiv

Data on arXiv recall-oriented understudy for gisting evaluation (ROUGE-1) was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on Papers With Code. The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (ROUGE-1) correspond to the result reported in the most recent version of each paper. Details about the arXiv benchmark are in the [arXiv dataset webpage](#).

To highlight progress on arXiv, scores were taken from the following papers:

[Big Bird: Transformers for Longer Sequences](#)

[A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents](#)

[Get to the Point: Summarization With Pointer-Generator Networks](#)

[Long Document Summarization With Top-Down and Bottom-Up Inference](#)

[MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes](#)

[PEGASUS: Pre-training With Extracted Gap-Sentences for Abstractive Summarization](#)

PubMed

Data on PubMed recall-oriented understudy for gisting evaluation (ROUGE-1) was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (ROUGE-1) correspond to the result reported in the most recent version of each paper. Details about the PubMed benchmark are in the [PubMed paper](#).

To highlight progress in PubMed, scores were taken from the following papers:

[A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents](#)

[Get to the Point: Summarization With Pointer-Generator Networks](#)

[Long Document Summarization With Top-Down and Bottom-Up Inference](#)

[LongT5: Efficient Text-to-Text Transformer for Long Sequences](#)

[PEGASUS: Pre-training With Extracted Gap-Sentences for Abstractive Summarization](#)

[Sparsifying Transformer Models With Trainable Representation Pooling](#)

Abductive Natural Language Inference (aNLI)

Data on Abductive Natural Language Inference (aNLI) was sourced from the Allen Institute for AI's [aNLI leaderboard](#). Details on the aNLI benchmark can be found in the [aNLI paper](#).



SST-5 Fine-Grained

Data on SST-5 Fine-Grained accuracy was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (accuracy) correspond to the result reported in the most recent version of each paper. Details about the SST-5 Fine-Grained benchmark can be found in the [SST paper](#).

To highlight progress on SST-5 Fine-Grained accuracy, scores were taken from the following papers:

[An Algorithm for Routing Capsules in All Domains](#)

[An Algorithm for Routing Vectors in Sequences](#)

[Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks](#)

[Improved Sentence Modeling Using Suffix Bidirectional LSTM](#)

[Learned in Translation: Contextualized Word Vectors](#)

[Less Grammar, More Features](#)

[Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank](#)

[Self-Explaining Structures Improve NLP Models](#)

MMLU

Data on MMLU accuracy was retrieved through a detailed arXiv literature review cross-referenced by technical progress reported on [Papers With Code](#). The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (accuracy) correspond to the result reported in the most recent version of each paper. Details about the MMLU benchmark can be found in the [MMLU paper](#).

To highlight progress on MMLU accuracy, scores were taken from the following papers:

[Language Models Are Few-Shot Learners](#)

[Language Models Are Unsupervised Multitask Learners](#)

[Scaling Instruction-Finetuned Language Models](#)

[Scaling Language Models: Methods, Analysis & Insights from Training Gopher](#)

Number of Commercially Available MT Systems

Details about the number of commercially available MT systems were sourced from the Intento report [The State of Machine Translation, 2022](#). Intento is a San Francisco-based startup that analyzes commercially available MT services.

VoxCeleb

Data on VoxCeleb equal error rate (EER) was retrieved from the [VoxCeleb Speaker Recognition Challenge \(VoxSRC\)](#).

For the sake of consistency, the AI Index reported scores on the initial VoxCeleb dataset. Specifically, the AI Index made use of the following sources of information:

[ID R&D System Description to VoxCeleb Speaker Recognition Challenge 2022](#)

[The IDLAB VoXSRC-20 Submission: Large Margin Fine-Tuning and Quality-Aware Score Calibration in DNN Based Speaker Verification](#)

[The SpeakIn System for VoxCeleb Speaker Recognition Challenge 2021](#)

[VoxCeleb: A Large-Scale Speaker Identification Dataset](#)

[VoxCeleb: Large-Scale Speaker Verification in the Wild](#)

[VoxCeleb2: Deep Speaker Recognition](#)

Whisper

Data on Whisper for large-scale speech recognition models was retrieved from the paper [Radford et al., 2022](#).

Procgen

Data on Procgen mean-normalized score was retrieved through a detailed arXiv literature review. The reported dates correspond to the year in which a paper was first published to arXiv, and the reported results (mean-normalized score) correspond to the result reported in the most recent version of each paper. Details on the Procgen benchmark can be found in the [Procgen paper](#).

To highlight progress on Procgen, scores were taken from the following papers:

[Automatic Data Augmentation for Generalization in Reinforcement Learning](#)

[Leveraging Procedural Generation to Benchmark Reinforcement Learning](#)

[Procedural Generalization by Planning With Self-Supervised World Models](#)

[Rethinking Value Function Learning for Generalization in Reinforcement Learning](#)

Training Time, Number of Accelerators, and Performance

Data on training time, number of accelerators, and performance for AI systems was taken from the MLPerf Training and Inference benchmark competitions. Details on the MLPerf Training benchmark can be found in the [MLPerf Training Benchmark paper](#), while details on MLPerf Inference can be found in the [MLPerf Inference Benchmark paper](#). Information about the current benchmark categories as well as technical information about submission and competition subdivisions can be found on the [MLPerf Training](#) and [MLPerf Inference](#) webpages.

The AI Index made use of data from the following MLPerf Training competitions:

[MLPerf Training v2.1, 2022](#)

[MLPerf Training v2.0, 2022](#)

[MLPerf Training v1.1, 2021](#)

[MLPerf Training v1.0, 2021](#)

[MLPerf Training v0.7, 2020](#)

[MLPerf Training v0.6, 2019](#)

[MLPerf Training v0.5, 2018](#)

The AI Index made use of data from the following MLPerf Inference competitions:

[MLPerf Inference v2.1, 2022](#)

[MLPerf Inference v2.0, 2022](#)

[MLPerf Inference v1.1, 2021](#)

[MLPerf Inference v1.0, 2021](#)

[MLPerf Inference v0.7, 2020](#)